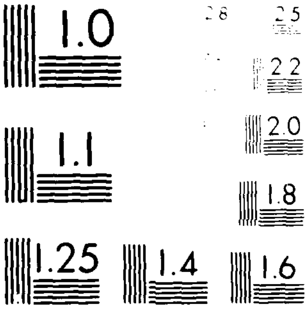
AD-A111 543　DECISIONS AND DESIGNS INC　MCLEAN VA　F/G 12/1
EVALUATING THE RELATIVE EFFECTIVENESS OF DIFFERENT STRUCTURING --ETC(U)
JAN 82　L ADELMAN, P J STICHA, M L DONNELL　N00014-81-C-0022
UNCLASSIFIED　TR-82-1-326-13　NL

1 OF 2
AD A
111 543


1.0
2.8
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

6

January 1982

TECHNICAL REPORT TR 82-1-326.13

Evaluating the Relative Effectiveness of Different Structuring and Weighting Techniques for Multi-Attribute Value Assessment

Leonard Adelman
Paul J. Sticha
Michael L. Donnell

DECISIONS and DESIGNS, INC.

Prepared for
Office of Naval Research
Psychological Sciences Division (Code 442EP)
800 North Quincy Street
Arlington, Virginia 22217

TECHNICAL REPORT TR 82-1-326.13

EVALUATING THE RELATIVE EFFECTIVENESS OF DIFFERENT STRUCTURING AND WEIGHTING TECHNIQUES FOR MULTI-ATTRIBUTE VALUE ASSESSMENT

by

Leonard Adelman, Paul J. Sticha, and Michael L. Donnell

Prepared for

Office of Naval Research
Psychological Sciences Division (Code 442EP)
800 North Quincy Street
Arlington, Virginia 22217
Contract N00014-81-C-0022

January 1982

DECISIONS and DESIGNS, INC.
Suite 600, 8400 Westpark Drive
P. O. Box 907
McLean, Virginia 22101
(703) 821-2828

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
TR 82-1-326.13

2. GOVT ACCESSION NO.
AD-A111543

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
EVALUATING THE RELATIVE EFFECTIVENESS OF DIFFERENT STRUCTURING AND WEIGHTING TECHNIQUES FOR MULTI-ATTRIBUTE VALUE ASSESSMENT

5. TYPE OF REPORT & PERIOD COVERED
Final technical report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Leonard Adelman
Paul J. Sticha
Michael L. Donnell

8. CONTRACT OR GRANT NUMBER(s)
N00014-81-C-0022

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Decisions and Designs, Inc.
Suite 600, 8400 Westpark Drive, P.O. Box 907
McLean, VA 22101

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR197069

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Psychological Sciences Division
800 N. Quincy Street, Arlington, VA

12. REPORT DATE
January 1982

13. NUMBER OF PAGES
100

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multi-attribute value assessment	Combat readiness
Multi-attribute structuring techniques	MCCRES
Multi-attribute weighting techniques	Marine Corps
Group discussion techniques	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three experiments were conducted to evaluate the relative effectiveness of different structuring and weighting techniques for Multi-Attribute Value Assessment (MAVA). In particular, the first experiment evaluated the relative effectiveness of two techniques for structuring the MAVA hierarchy; the second experiment evaluated the relative effectiveness of five techniques for obtaining an individual's weights on attributes in the hierarchy; and the third experiment evaluated the relative effectiveness of two weighting techniques,

in conjunction with two discussion techniques, for obtaining group weights on attributes in the hierarchy. In all three experiments, the participants were second lieutenants in the U.S. Marine Corps who had completed their training at The Basic School. The external criterion was the MAVA model for the ATTACK Mission Performance Standard (MPS) in the Marine Corps Combat Readiness Evaluation System (MCCRES).

The thesis guiding all three experiments was that the effectiveness of different structuring and weighting techniques depends on task characteristics; this was confirmed for weighting techniques, but not structuring techniques. This report provides a detailed technical description of the three experiments. In addition, some brief comments regarding the potential use of MAVA techniques as diagnostic tools for identifying the relative strengths and weaknesses of training programs are provided for the more operationally-oriented reader.

CONTENTS

CONTENTS (Cont'd)

FIGURES

TABLES

SUMMARY

There is a paucity of systematic research that evaluates the relative effectiveness of different multi-attribute value and utility assessment techniques for tasks representative of actual decision problems and for which an external criterion exists. This paucity exists because such tasks are rare. Yet, such research is required if decision analysts are to learn which multi-attribute value and utility assessment techniques are most effective for actual decision problems. Decisions and Designs, Inc. (DDI) was given a unique opportunity to perform such research as a result of previous success in developing a Multi-Attribute Value Assessment (MAVA) model to measure combat readiness; the model and its supporting materials is called the Marine Corps Combat Readiness Evaluation System (MCCRES). This report describes the technical work performed by DDI analysts. Some brief comments regarding the potential use of MAVA techniques as diagnostic tools for identifying the relative strengths and weaknesses of training programs are also provided for the more operationally-oriented reader.

Three experiments were conducted to evaluate the relative effectiveness of different structuring and weighting techniques for MAVA. The first experiment evaluated the relative effectiveness of two techniques for structuring the MAVA hierarchy; the second experiment evaluated the relative effectiveness of five techniques for obtaining an individual's weights on attributes in the hierarchy, and the third experiment evaluated the relative effectiveness of two weighting techniques, in conjunction with two discussion techniques, for obtaining group weights on attributes in the hierarchy. The guiding thesis throughout all three experiments was that the relative effectiveness of different techniques depends on task characteristics.

The participants in all three experiments were second lieutenants in the Marine Corps, who had just completed the Infantry Officers Course and, consequently, their formal training at The Basic School. The external criterion was the MAVA model for the ATTACK Mission Performance Standard (MPS) in MCCRES. It is believed that this represents the first setting representative of an actual decision-making problem where an existing external criterion has been available for evaluating the effectiveness of both structuring and weighting techniques.

Top-Down and Bottom-Up structuring approaches were evaluated in the first experiment. There were three classes of dependent variables: the relative effectiveness of the two approaches in identifying bottom-level nodes in the ATTACK MPS; the shape of the hierarchies generated by the approaches; and the participants' and analysts' subjective ratings of the quality of the group's hierarchy and discussion. The first hypothesis was that the Top-Down approach would be more effective than the Bottom-Up approach because it was more congruent with the formal characteristics of the task. The results did not support this hypothesis; there was no difference in the relative effectiveness of the two structuring approaches. The second hypothesis was that the Top-Down approach would result in fewer top-level nodes, a deeper hierarchy, more terminal nodes, a smaller average node width and a larger ratio of nonterminal to terminal nodes than the Bottom-Up approach. The results supported the last two points, and found that the mean depth of the single deepest branch was larger for the Top-Down approach. Third, no hypotheses were made regarding the subjective ratings. Although there were no significant differences in either the participants' or the analysts' subjective ratings of the two approaches, the participants were generally more satisfied than the analysts with the group's hierarchy and discussion.

Experiment 2 evaluated the relative effectiveness of five different techniques for obtaining an individual's weights on attributes in four tasks in the ATTACK MPS. The five weighting techniques were (1) Edward's (1977) ratio estimation technique for riskless choice (called SMART for Simple Multi-Attribute Rating Technique); (2) a paired comparison technique that forces participants to explicitly make the comparisons implicit in Edward's technique; (3) Hammond et al.'s (1975) statistical decomposition technique (uses multiple regression analysis) called "policy-capturing"; (4) Keeney and Raiffa's (1976) lottery technique for risky choice; and (5) a simple, subjective rating technique, that of dividing up 100 points among the attributes. The four tasks varied along two dimensions: the number of attributes (either 5 or 9) and the peakness of the distribution of actual task weights (either peaked or flat). It was hypothesized that the smaller the number of attributes, the more effective the weighting techniques because of fewer information processing requirements. It was further hypothesized that the relative effectiveness of different weighting techniques would depend on the distribution of task weights because different techniques systematically affect the standard deviation (spread) of individual participants' weights. Both hypotheses were confirmed.

The third experiment evaluated the relative effectiveness of two weighting techniques, in conjunction with two discussion techniques, for obtaining a group's weights on the attributes in the four tasks used in Experiment 2. SMART and policy-capturing were the weighting techniques; the Nominal Group Technique (NGT) and a leader-directed technique were the discussion techniques. It was hypothesized that the standard deviations of a group's weights was a mediating variable of their relative effectiveness. This hypothesis was not confirmed, which was surprising for two reasons. First, a replication study again found that the standard deviations were a mediating variable of the relative effectiveness of individual participants' weights. SMART led to a much larger standard deviation of individual participants' weights than did policy-capturing and, as in Experiment 2, greater accuracy when the distribution of task weights was peaked and lower accuracy when the distribution was flat. Second, all of the hypotheses were confirmed for the standard deviations of the group weights. The leader-directed discussion technique led to a larger standard deviation of the group weights than did the NGT, and SMART led to a larger standard deviation than policy capturing. Yet, the predicted discussion by distribution, weighting by distribution, and discussion by weighting by distribution interactions were not significant for the accuracy of the group weights. The only significant findings for group accuracy were a main effect for the number of attributes and a number by distribution interaction. Accuracy was significantly higher for the five than for the nine attribute task. When the task had five attributes, accuracy was higher for the task with a peaked distribution; when the task had nine attributes, accuracy was higher for the task with a flat distribution.

In summary, the thesis guiding the research reported herein was that the task characteristics would affect the relative effectiveness of different structuring and weighting techniques. This thesis was strongly confirmed for weighting techniques, but not for structuring techniques. Future research should expand the investigation to substantive and formal task characteristics not considered in the present experiments. Specific recommendations for future research are presented separately in the discussion section for each of the three experiments.

The final section of the report discusses the potential use of MAVA techniques as diagnostic tools for identifying the relative strengths and weaknesses of Marine Corps training programs, or more generally, any training program for which an external criterion exists. Experiment 1, for example, illustrated how multi-attribute structuring techniques could be used as an effective diagnostic tool. Although the performance of second lieutenants participating in the first session of the experiment met Marine Corps standards, they did not identify as many requirements as our evaluator and their instructor thought they should have for the Consolidation task in the ATTACK MPS. The curriculum for teaching the ATTACK within the Infantry Officers Course was changed to place more emphasis on the Consolidation phase as a result of this finding. Other curriculum changes also occurred as a result of subsequent sessions, clearly demonstrating that Marine Corps training personnel found the multi-attribute structuring techniques to be effective diagnostic tools.

ACKNOWLEDGMENTS

The authors wish to thank Dennis M. Buede, Gary A. Frisvold, Kenneth P. Kuskey, Anne W. Martin, Cameron R. Peterson, Jonathan J. Weiss, all of Decisions and Designs, Inc.; Lieutenant Colonel Paul R. Catalogne, USMC; Lieutenant Colonel Ron K. Roth, USMC; as well as personnel at The Basic School, Quantico Marine Base, Virginia, for their assistance throughout the course of the project.

EVALUATING THE RELATIVE EFFECTIVENESS OF DIFFERENT STRUCTURING AND WEIGHTING TECHNIQUES FOR MULTI-ATTRIBUTE VALUE ASSESSMENT

1.0 INTRODUCTION

There is a paucity of systematic research evaluating the relative effectiveness of different multi-attribute value and utility assessment techniques for tasks representative of actual decision problems and for which an external criterion exists. This paucity exists because, quite simply, such tasks are rare. Yet, such research is extremely important, for multi-attribute value and utility assessment techniques are routinely used in actual settings. Decisions and Designs, Inc. (DDI), for example, has applied such techniques for the last ten years to problems faced by a number of different agencies within the Departments of Defense, State, Transportation, Commerce, and Energy. Research with tasks representative of actual decision problems and for which an external criterion exists is required if decision researchers and analysts are to learn what multi-attribute value and utility assessment techniques are most effective for actual decision problems.

DDI had a unique opportunity to perform such research as a result of previous success in developing a Multi-Attribute Value Assessment (MAVA) model for measuring combat readiness. In 1976-1977, the Defense Advanced Research Projects Agency (DARPA) funded an exploratory effort by DDI to assist the U.S. Marine Corps in developing a sound methodological foundation for evaluating the readiness of combat units. DDI chose Multi-Attribute Value Assessment (MAVA) as the methodological foundation for this exploratory effort. Value assessment was selected instead of utility assessment because the five participating Marine Corps officers did not want to explicitly deal with uncertainty; consequently, there was no need to assess utility functions to represent their risk preference. The MAVA model developed by the working group was successfully field-tested by the Marine Corps in August of 1977. The model and supporting materials are now the standard combat readiness method for the Marine Corps. The method is called MCCRES, for Marine Corps Combat Readiness Evaluation System.

DDI has continued to work with Marine Corps personnel tasked with improving combat readiness. The research reported herein was conducted with the assistance of personnel in the Operations Division of the Plans, Policies, and Operations Department at Marine Corps Headquarters and with personnel teaching the Infantry Officers Course at the Basic School. Using the MAVA model within MCCRES as an external criterion, and working with Marine Corps second lieutenants who had just completed the Infantry Officers Course at The Basic School, three experiments were performed to evaluate the relative effectiveness of different structuring and weighting techniques for MAVA. It is believed that this is the first setting representative of an actual decision-making problem where an existing external criterion has been available to evaluate the effectiveness of both structuring and weighting techniques.

The first experiment evaluated the relative effectiveness of two techniques for structuring the MAVA hierarchy; the second experiment evaluated the relative effectiveness of five techniques for obtaining an individual's weights on attributes in the hierarchy; and the third experiment evaluated the relative effectiveness of two discussion techniques, for obtaining group weights on attributes in the hierarchy. The guiding thesis throughout all three experiments was that the relative effectiveness of different decision-analytic techniques depends on task characteristics. This theoretical perspective is a Brunswikian one (for a general presentation, see Brunswik, 1955; Hammond, 1966), which at its most general level argues that organismic behavior and subsequent performance is significantly, and at times solely determined by environmental characteristics. Hammond (1981) and his colleagues (e.g., see Hammond et al., 1975; Hammond and Wascoe, 1980) have adapted this perspective to the study of human judgment and decision making by studying how judgment is affected by different formal and substantive characteristics of the judgment task. Although Hammond and his colleagues have never studied how task characteristics affect the judgmental accuracy achieved with different MAVA techniques, the focus of the three experiments reported herein and their theoretical perspective, has guided the present research effort.

This report provides a detailed technical description of each of the three experiments. The concluding section presents some brief comments for the more operationally-oriented reader regarding the potential use of MAVA techniques for identifying the relative strengths and weakness of training programs.

2.0 EXPERIMENT 1: STRUCTURING THE MULTI-ATTRIBUTE HIERARCHY

2.1 Introduction

Structuring the Multi-Attribute Value Assessment (MAVA) problem is the most important phase of MAVA because the outcome is dependent on it. For example, Von Winterfeldt and Edwards (1975) did a simulation study identifying the ease with which dominated alternatives can be selected through the improper structuring of the problem. Yet, we know "... next to nothing about eliciting the structure of problems from decision makers." (Fischhoff, 1977, p. 10). Research has not been done on structuring the multi-attribute problem in settings with an accepted external criterion.

Few structuring techniques have been proposed by decision analysts. The following two multi-attribute structuring techniques are, however, good candidates for experimentation because they are used by practicing decision analysts: a hierarchical (or Top-Down) approach (Raiffa, 1969; Keeney and Raiffa, 1976) and DDI's attribute-listing (or Bottom-Up) approach. The Top-Down approach to structuring the problem proceeds exactly like the multi-attribute hierarchy looks on paper; the upper-level nodes are listed first and then each node, in turn, is subdivided into its component attributes until the process identifies the lowest-level attributes. In contrast, the Bottom-Up approach to structuring the problem proceeds by first having participants list all of the possible characteristics (i.e., lower-level nodes) of the hierarchy without any concern for the hierarchical nature of the problem; attributes are subsequently clustered together to form the multi-level branches moving up the hierarchy.

2.2 Method

The Top-Down and Bottom-Up structuring techniques were evaluated experimentally. The twenty USMC second lieutenants participating during each session were divided into four groups of five participants each. One trained DDI analyst led each group. The analyst used one of the two structuring techniques to help the group develop a multi-attribute hierarchy for the ATTACK Mission Performance Standard (MPS) in MCCRES (Figure 2-1). Two of the analysts used the Top-Down technique; two used the Bottom-Up technique. Approximately three hours (from 9:00 a.m. to noon) were spent developing the hierarchy. The group then spent approximately one more hour examining the correct ATTACK MPS and identifying which bottom-level nodes in that MPS were in their structure. (Appendix A contains the analysts' instructions for each structuring technique.)

Seven sessions were conducted during the study. In all, six analysts participated; two of the analysts used the Top-Down technique for three groups and the Bottom-Up technique for three groups. The other four analysts used each technique twice. All participating second lieutenants had just completed their training in the Infantry Officers Course at the Basic School.

There were three types of dependent variables. The first type was the relative effectiveness of the two structuring techniques in identifying bottom-level nodes in MCCRES. It was hypothesized that the Top-Down approach would be more effective than the Bottom-Up approach because it was more congruent with the formal task characteristics of the problem under consideration. The ATTACK MPS is a very systematic, almost analytical decomposition problem; a number of distinct tasks must be performed, and each task has distinct performance

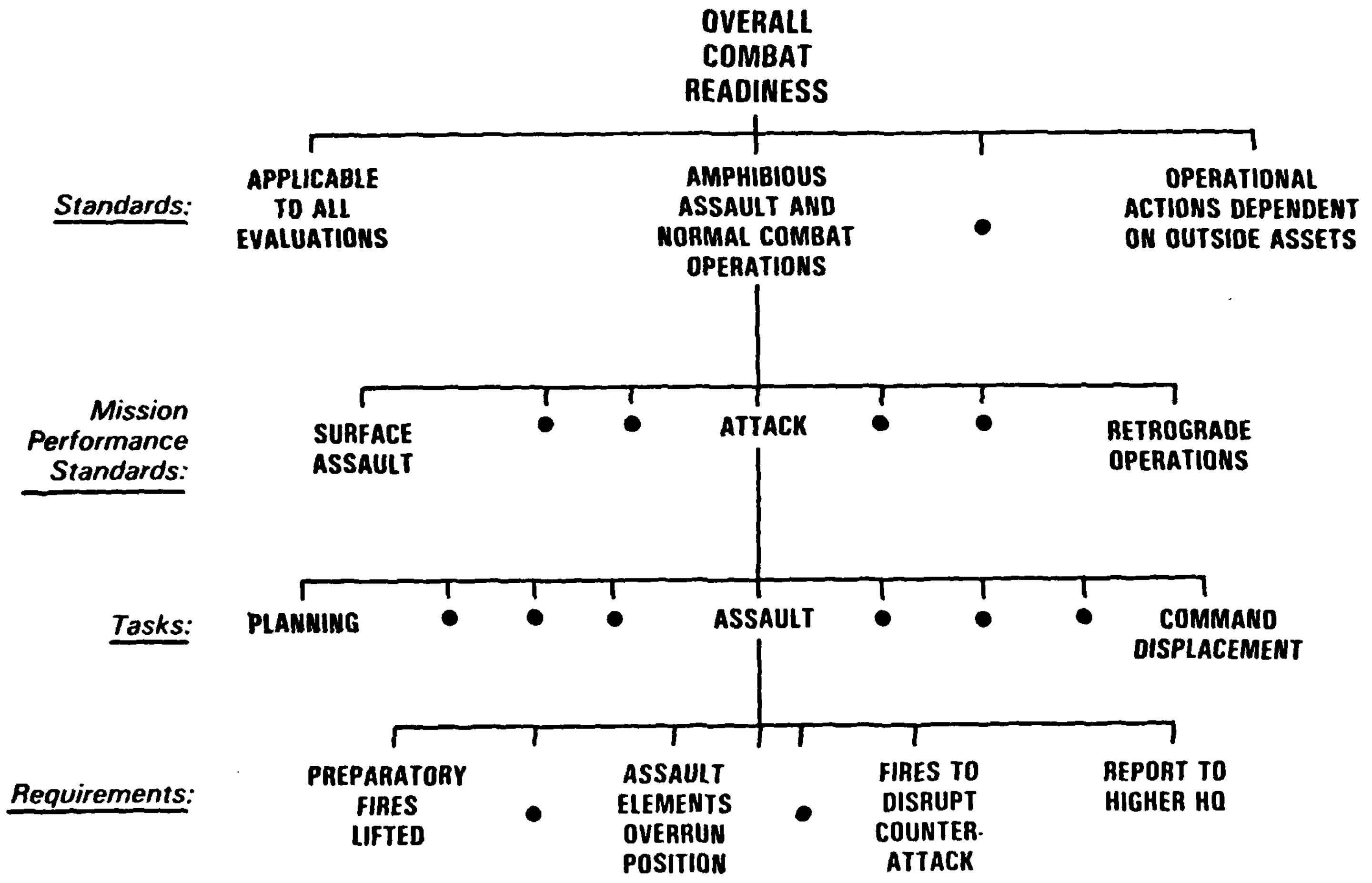

OVERALL COMBAT READINESS
Standards:
APPLICABLE TO ALL EVALUATIONS
AMPHIBIOUS ASSAULT AND NORMAL COMBAT OPERATIONS
●
OPERATIONAL ACTIONS DEPENDENT ON OUTSIDE ASSETS
Mission Performance Standards:
SURFACE ASSAULT
● ●
ATTACK
● ●
RETROGRADE OPERATIONS
Tasks:
PLANNING
● ● ●
ASSAULT
● ● ●
COMMAND DISPLACEMENT
Requirements:
PREPARATORY FIRES LIFTED
●
ASSAULT ELEMENTS OVERRUN POSITION
●
FIRES TO DISRUPT COUNTER-ATTACK
REPORT TO HIGHER HQ

Figure 2-1

PICTORIAL REPRESENTATION OF THE MAV HIERARCHY FOR THE ATTACK MPS IN MCCRES requirements. The Top-Down approach is a much more systematic structuring approach than the Bottom-Up approach; the former explicitly decomposes a problem from its most general to most specific components, while the latter does not. Instead, the Bottom-Up approach focuses on the differences between the alternatives available to the decision maker in describing the specific bottom-level attributes. Since there were no well-defined alternatives available for MCCRES structuring, it was hypothesized that the relative effectiveness of the Bottom-Up approach would be reduced for the ATTACK MPS. In summary, the Top-Down approach better matches the formal task requirements of the ATTACK MPS and, therefore, was hypothesized to be more effective than the Bottom-Up approach in identifying bottom-level nodes.

The measures of relative effectiveness were determined by the number of correctly identified requirements (i.e., terminal nodes) in the ATTACK MPS, in the opinion of the participating second lieutenants and an independent rater. The lieutenant colonel in the Operations Division of the Plans, Policies, and Operations Department at Marine Corps Headquarters, who participated in writing the MCCRES volume for infantry units, was the independent rater. He evaluated the MAVA hierarchy developed by each group within six weeks after the session with the participating second lieutenants. When performing this task, he was, of course, not aware of which technique was used to generate each group hierarchy.

The second type of dependent variable was the shape (or configuration) of the MAVA hierarchies generated by each technique. It was hypothesized that the Top-Down approach would result in hierarchies that looked considerably different than the ones generated by the Bottom-Up approach. In particular, it was hypothesized that the Top-Down approach would result in

(1) fewer top-level nodes, (2) a deeper hierarchy, (3) more terminal nodes, (4) a smaller average node width, and (5) a larger ratio of non-terminal to terminal nodes than the Bottom-Up approach. These hypotheses were based on the experience of DDI analysts who have used both structuring techniques. They represent hypothesized differences in the implications on human information processing of using each structuring technique. In particular, DDI analysts hypothesized that the Top-Down approach generally would result in an in-depth decomposition of the major aspects of the problem under consideration. In contrast, it was hypothesized that the Bottom-Up approach generally would result in a less detailed, but broader decomposition of the problem. These differences are represented in Figure 2-2.

The third type of dependent variable was the subjective ratings of both the analyst and the participating second lieutenants as to the quality of the group's MAVA structure and characteristics of their discussion. In particular, the analysts answered questions regarding the following seven topics listed below, using a seven-point Likert scale ranging from very low (#1) to very high (#7):

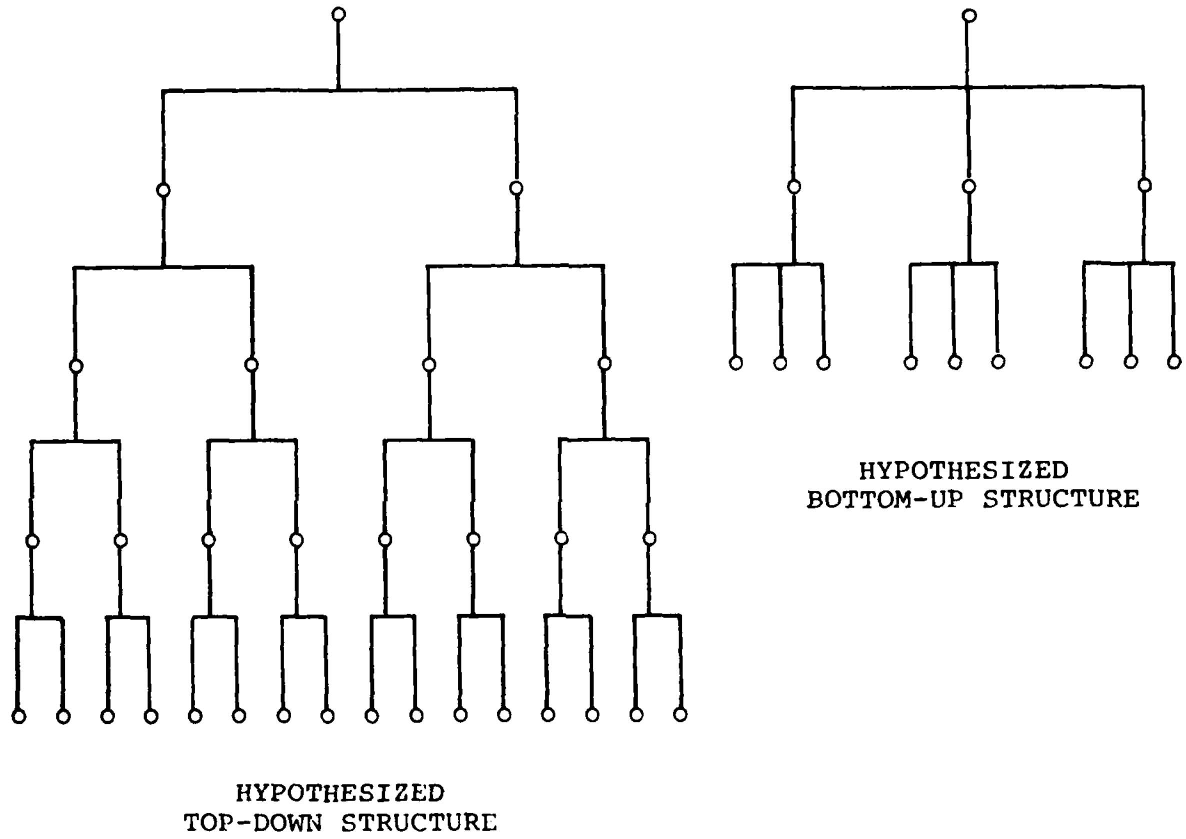
1. Satisfaction with the group's structure.

2. Quality of the group's discussion.

3. Extent to which the group's discussion involved the repetition of ideas and suggestions.

4. Frequency with which each person was able to speak.

5. Frequency with which voices were raised during the group's discussion.



HYPOTHESIZED
TOP-DOWN STRUCTURE

HYPOTHESIZED
BOTTOM-UP STRUCTURE

Figure 2-2

PICTORIAL REPRESENTATION OF HYPOTHESES REGARDING THE SHAPE OF THE HIERARCHY FOR DIFFERENT STRUCTURING TECHNIQUES

6. Extent to which the group's discussion was directed by, or centered around, one group member.

7. Extent to which the analysts were able to implement the specified structuring approach.

USMC group members only answered the first six questions. As far as it is known, no study has investigated differences in the opinions of analysts and group members as to the quality of the analysis and discussion. Yet, the success or failure of decision-analytic efforts on applied problems primarily depends on the satisfaction and subsequent support of the participants because external criteria for measuring accuracy, seldom exist outside a laboratory setting. We had no basis for hypothesizing any differences between the two techniques' mean scores for the first six subjective ratings; however, we did hypothesize that analysts would be better able to implement the Top-Down approach because it matches the formal task characteristic of the ATTACK MPS better than the Bottom-Up approach.

2.3 Results

T-tests were performed to determine whether the two approaches led to significant differences in the mean number of correctly identified MCCRES requirements (i.e., terminal nodes in the ATTACK MPS). A $p < .05$ level of significance was used for a two-tailed test. According to the assessments made by both the independent rater and the participants, there were no differences in the mean number of requirements correctly identified by the Top-Down and Bottom-Up approaches. According to the independent rater, for example, $\bar{x}_{TD} = 30.36$, $\bar{x}_{BU} = 27.79$, $t = 1.17$, $df = 26$, $p > .10$.

According to the assessments of the independent rater alone, the participating second lieutenants met performance standards for all nine tasks. The groups identified, on an average, 44% of the requirements in the ATTACK MPS. On the average, these assessments accounted for 48% of the possible cumulative utility for this criterion. Performance was significantly better for the first six tasks in the ATTACK MPS, which are the tasks in this MPS for which USMC second lieutenants receive more training. On the average, participants identified 57% of the requirements and 62% of the possible cumulative utility for these six tasks. Furthermore, they almost always identified the most important requirement for each of these tasks, obtaining 94% of the cumulative utility for them in the ATTACK MPS.

Two-tailed t-tests were performed to test the hypothesized differences in the shape of the hierarchies generated by the two approaches. The results of these tests are presented in turn.

1. There was no difference in the mean number of top-level nodes (or branches): $\bar{x}_{TD}=4.79$, $\bar{x}_{BU}=4.43$, $t = .42$, $df = 26$, $p > .10$.

2. The mean depth of the single deepest branch was significantly larger for the Top-Down approach ($\bar{x}_{TD} = 4.71$, $\bar{x}_{BU} = 4.07$): $t = 2.34$, $df = 26$, $p < .05$. There was, however, no significant difference in the mean depth across all terminal nodes: ($\bar{x}_{TD} = 3.27$, $\bar{x}_{BU} = 3.0$); $t = 1.46$, $df = 26$, $p > .10$.

3. There was no significant difference in the mean number of terminal nodes: $\bar{x}_{TD} = 67.42$, $\bar{x}_{BU}= 60.29$, $t = 1.06$, $df = 26$, $p >.10$.

4. The average node width was measured by equation [1]:

$$\text{Average Node Width} = \frac{\text{Total \# Nodes}}{\text{Total Nodes} - \text{\# Terminal Nodes} + 1} \qquad [1]$$

The +1 was placed in the denominator because the top-level node was not included in the total number of nodes. The average node width for the Top-Down approach was significantly smaller than for the Bottom-Up approach ($\bar{x}_{TD} = 3.68$, $\bar{x}_{BU} = 4.65$); $t = 2.4$, $df = 26$, $p < .05$.

5. The ratio of nonterminal to terminal nodes was measured by equation [2]:

$$\frac{\text{\#Nonterminal Nodes}}{\text{\#Terminal Nodes}} = \frac{(\text{Total \# Nodes} - \text{\# Terminal Nodes}) - 1}{\text{\# Terminal Nodes} - 2} \qquad [2]$$

As can be seen in Figure 2-3, this measure has a range between 0.0 and 1.0, with 0.0 representing a hierarchy with only one level of decomposition and 1.0 representing a hierarchy where each node is divided into only two lower-level nodes, the smallest degree of decomposition. Although this measure is related to the average node width, examples 2 and 3 in Figure 2-3 illustrate that two hierarchies can have the same average node width, but have different ratios of nonterminal to terminal nodes. In the present study, as hypothesized, the ratio of nonterminal to terminal nodes was significantly larger for the Top-Down approach ($\bar{x}_{TD} = .37$, $\bar{x}_{BU} = .28$); $t = 2.39$, $df = 26$, $p < .05$.

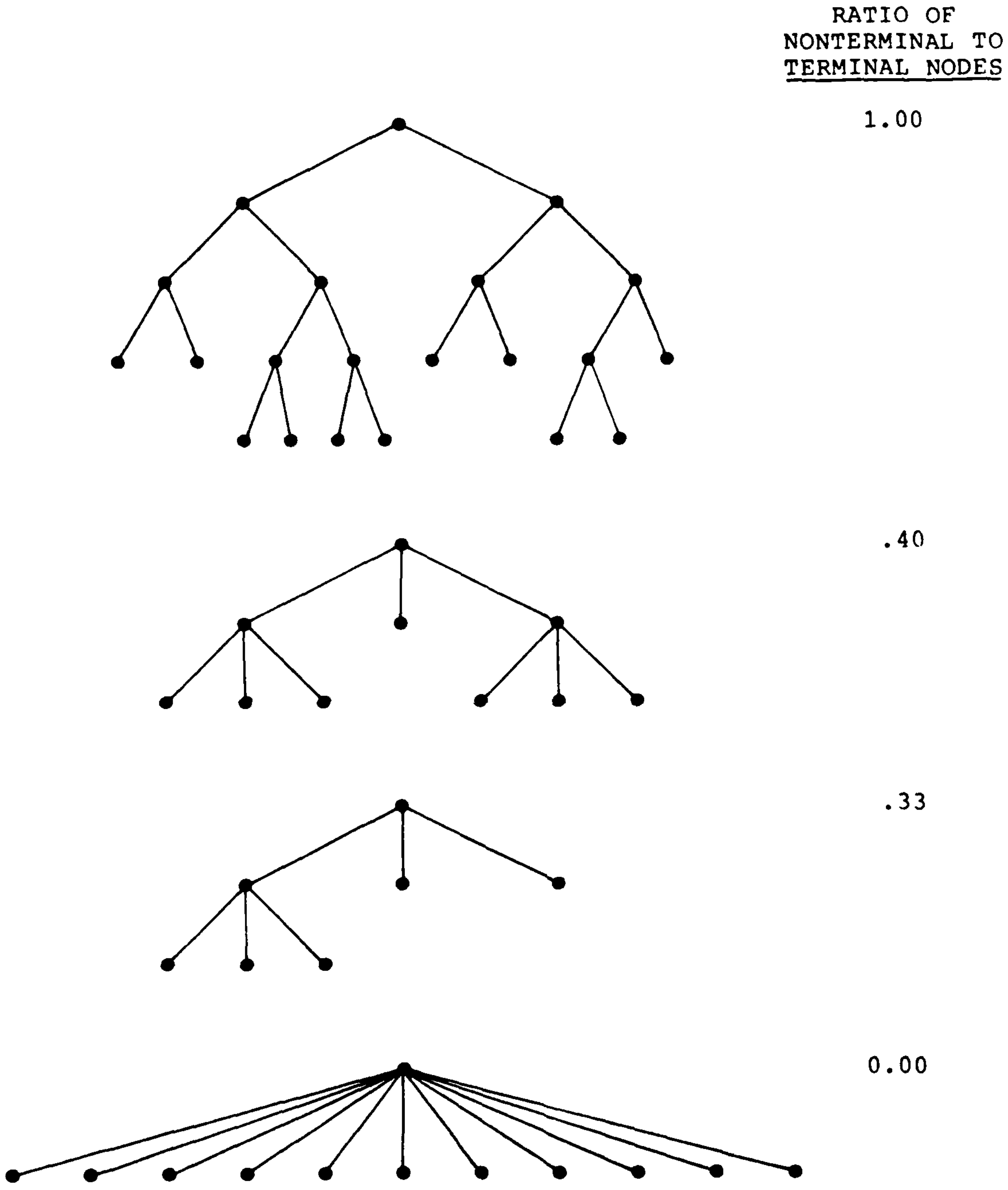


Figure 2-3

AN EXAMPLE OF HIERARCHIES WITH DIFFERENT RATIOS OF NONTERMINAL TO TERMINAL NODES

In summary, significant differences were found in the mean shape of the hierarchies generated by th^ two structuring techniques; the study, however, did not confirm all the hypotheses.

There were no significant differences in either the participants' or the analysts' subjective ratings for the two techniques. The only subjective measure that approached significance was the analysts' rating for implementability ($\bar{x}_{TD} = 5.50$, $\bar{x}_{BU} = 4.65$: $t = 1.87$, $df = 18$ because assessment did not commence until the third session, $.05 < p < .10$ for a two-tailed test). The direction of this difference corresponds to an assessment that the Top-Down approach was easier to implement than the Bottom-Up approach.

There were, however, significant differences between the subjective ratings of participants and analysts. The unit of analysis for the participants' ratings was the mean score for each five-person group. Two-tailed t-tests were performed between the mean group ratings and the analysts' ratings for each subjective measure. As can be seen in Table 2-1, participants were significantly more satisfied than the analysts with the group's hierarchy and discussion. Although participants thought there was signficantly greater participation than did analysts, they also thought there was greater repetition of ideas and more occasions during which voices were raised. This finding might suggest a response bias. The only rating where there were no significant differences between participants and analysts was on the extent to which the discussion was directed by one person.

SUBJECTIVE MEASURES	PARTICIPANTS' MEAN RATINGS	ANALYSTS MEAN RATINGS
1. Satisfaction with structure	5.84	4.87**
2. Quality of discussion	5.80	5.13**
3. Repetition of ideas	4.22	2.83***
4. Extent of participation	5.96	5.38*
5. Extent to which voices raised	2.68	1.71*
6. Discussion directed by one person	3.01	2.73

* $p < .05$, df = 46, 2-tailed test
** $p < .01$, df = 46, 2-tailed test
*** $p < .001$, df = 46, 2-tailed test

Table 2-1

PARTICIPANTS AND ANALYSTS MEAN RATINGS
FOR SUBJECTIVE MEASURES

2.4 Discussion

The study reported herein is believed to be the first empirical effort to evaluate the relative effectiveness of different techniques for structuring the MAVA hierarchy. It was hypothesized that the Top-Down approach would lead to a more accurate group hierarchy than the Bottom-Up approach. This hypothesis rested on the assumption that the Top-Down approach was more congruent with the formal task characteristics of the ATTACK MPS than was the Bottom-Up approach. The ATTACK MPS is a very systematic, almost analytical decomposition problem; a number of distinct tasks must be performed in a sequential fashion, and each task has distinct performance requirements. Since the Top-Down approach is more analytically structured than the Bottom-Up approach, it is more congruent with the task characteristics of the ATTACK MPS. Consequently, it was hypothesized that it would be better able to generate MAVA hierarchies containing the terminal nodes (requirements) in the ATTACK MPS. The experiment did not support this hypothesis.

On the basis of these results, there is no support for the proposition that the relative effectiveness of different structuring techniques depends on the congruence between technique and task characteristics. A post hoc analysis, however, suggests an alternative explanation for these results that is consistent with the proposition; the experiment did support, to some extent, the hypothesis that the shape of the MAVA hierarchy would be dependent on the structuring technique. As hypothesized, the Top-Down approach resulted in hierarchies with a smaller average node width and a larger ratio of nonterminal to terminal nodes than those generated by the Bottom-Up approach. A post hoc analysis found the average node width of the MAVA hierarchy in the ATTACK MPS to be 7.5; this is closer to the node width obtained with the Bottom-Up approach ($\bar{x}_{BU}$ = 4.65) than with the Top-Down approach ($\bar{x}_{TD}$ = 3.68). Furthermore, the ratio of nonterminal to terminal nodes in the ATTACK MPS was .125, which is significantly closer to the figure for the Bottom-Up approach ($\bar{x}_{BU}$ = .28) than the figure for the Top-Down approach ($\bar{x}_{TD}$ = .37). It is, therefore, admissible to argue that the reason there were no significant differences in the relative effectiveness of the two approaches was because the two approaches were congruent, but in different ways. The Top-Down approach was more congruent with the analytical decomposition format of evaluation systems like the one represented in the ATTACK MPS; in contrast, the Bottom-Up approach was more consistent with the shape of the MAVA hierarchy in the ATTACK MPS.

This alternative explanation is post hoc and obviously, should be taken with caution. It can be tested in future research, for example, by selecting a task with both an analytical decomposition format and a MAVA hierarchy with a small average node width and a large ratio of nonterminal to terminal nodes. In this way, the proposition that the relative effectiveness of different structuring techniques depends on task characteristics provides a framework for guiding future structuring research. Failure to find significant differences in the relative effectiveness of different structuring techniques over variation in the task characteristics thought to elicit these differences, would quickly and cost-efficiently confirm the results reported herein.

Future research also should investigate the participants' general level of accuracy in identifying the (bottom-level) attributes in the hierarchy; as far as it is known, the present study is the first one to do so. Three recently conducted studies had suggested that accuracy might be quite poor. In studying fault trees, which have a structural representation very similar to multi-attributed hierarchies, Fischhoff, Slovic, and Lichtenstein (1978) found that expert and non-expert participants were insensitive to omissions in the tree's structure, and that this insensitivity was not easily susceptible to modification. In a study of act (or option) generation, Pitz, Sachs, and Heerboth (1980) observed that, on the average, individual participants did not seem to generate a very complete set of acts, averaging less than a third of the acts the experimenters thought were "worth considering." In addition, Gettys, Manning, and Casey (1981) found that, for the most part, individual participants were unable to generate high utility acts, obtaining only between 20% to 30% (depending on the problem and analysis) of the cumulative utility for them.

In comparison to these results, the groups participating in the present experiment performed quite well. According to the assessments of the independent rater, the groups identified, on the average, 44% of the requirements in the ATTACK MPS. In general, these assessments accounted for 48% of the possible cumulative utility for this criterion. Although standards were met for all nine tasks, performance was significantly better for the six tasks for which the participants had received more training. Participants identified (on an average) 57% of the requirements and 62% of the possible cumulative utility for these six tasks. Furthermore, they almost always identified the most important requirement for each of these tasks, obtaining 94% of the cumulative utility for them in the ATTACK MPS.

The present study differed from the three, referenced above, in four significant ways. First, there was a widely accepted and widely used external criterion for accuracy in the present study. The criterion in the other three studies was developed by the researchers. Second, the present study evaluated participants' accuracy in identifying bottom-level attributes in a multi-attributed hierarchy. It may be that individuals are better at identifying evaluation criteria than generating options, although research on identifying attributes in fault trees (Fischhoff et al. (1978)) suggests that this may not be so. Third, the present study used participants trained in the substantive material employed in develcping the hierarchy; only Fischhoff et al. did so. Participants in the present study performed considerably better than did Fischhoff et al.'s, although procedural differences in the two studies make the comparison difficult. Fourth, participants in the present study worked in groups and were led by a trained decision analyst; participants worked individually in the other studies. Research reviewed by Delbecq, Van de Van, and Gustafson (1975) on brainstorming and other group techniques suggests that group interaction facilitates performance by aggregating different aspects of the problem focused on by different individuals. These four differences should be kept in mind when designing future research in order to achieve a better understanding of which factors facilitate and inhibit participants' accuracy when structuring the multi-attribute problem.

3.0 EXPERIMENT 2: OBTAINING AN INDIVIDUAL'S WEIGHTS FOR MULTIPLE ATTRIBUTES

3.1 Introduction

John and Edwards (1978) reviewed studies comparing the relative effectiveness of the two general approaches for obtaining an individual's weights on multiple attributes: "direct subjective estimation and indirect holistic estimation" (p. i). The direct subjective estimation techniques included ranking, fractionation, subjective-estimative methods, and paired comparisons. Indirect holistic techniques included unbiased and biased regression analysis, the ANOVA and fractional ANOVA paradigms, and the indifference techniques of pricing out and trading off to the most important dimension. John and Edwards (1978, p. 48) concluded:

> For many of the laboratory and field settings studied, subjects gave responses to direct subjective assessments of importance weights that were both consistent (high convergent validity) and accurate (high criterion validity). Few discrepancies were observed in studies comparing direct subjective estimates of importance to statistical indices of importance derived indirectly from holistic evaluations.

More recent studies by John and Edwards (1978b), John, Edwards, and Collins (1980), and Stillwell, Barron, and Edwards (1980) support this conclusion, although decomposition procedures did do somewhat better than the one holistic procedure used in the last study.

It was hypothesized that this conclusion was premature because the relative effectiveness of different weighting techniques has not been studied over systematic variation in task characteristics. Research comparing the relative effectiveness of equal with differential weights has demonstrated that task properties are important predictors of performance. Einhorn and Hogarth (1975) and Wainer (1976), for example, have shown that deviations from optimal weighting do not make a practical difference in the overall correlations between predictions and actual scores when one knows the signs of the predictor variables and either there is (a) low task predictability (e.g., R .5), (b) a small sample size (e.g., n 50), and a large number of predictors (e.g., k 8), or (c) a moderate-to-high positive correlation between the predictors (e.g., r .5). Regarding this last point, Newman (1977), McClelland (1978), and Stillwell, Seaver, and Edwards (1981) have demonstrated that the correlation between predictors is typically negative in decision problems where one must select one of a restricted number of alternatives; under this condition, differences in weights do make a difference in accuracy. In conclusion, the research clearly demonstrates that task properties determine the relative effectiveness of equal and differential weights.

It was believed that this finding would generalize to different multi-attribute weighting techniques. In particular, it was hypothesized that the number of task attributes and the peakness of the distribution of "true" task weights were two task characteristics that would affect the relative effectiveness of different weighting techniques. Fischer (1977), for example, found a high correlation between Raiffa and Keeney's (1976) riskless and risky decomposition techniques. Von Winterfeldt and Edwards (1975), however, found a substantially lower correlation between riskless and risky decomposition techniques. Fischer (1979) hypothesized that this discrepancy was due to the number of attributes used in the two studies; Fischer (1977) used three, while Von Winterfeldt d Edwards (1975) used fourteen. A study by Cook and Stewart (1975) provides tentative support for Fischer's hypothesis because they found that the correlations between the predicted values generated by seven different subjective weighting techniques and participants' actual values decreased as the number of attributes increased in the task; although not tested, the implication is that accuracy would have decreased too. Taken together, these findings suggest that all weighting techniques will be more effective the smaller the number of attributes, because less information processing is required.

It was also hypothesized that the relative effectiveness of different weighting techniques would depend on the distribution of task weights. Some weighting techniques (e.g., paired comparisons) tend to generate a peaked distribution of weights because they force individuals to rigorously compare the worth of one attribute relative to another, thereby spreading their weights. Other weighting techniques (e.g., a lottery technique) do not force individuals to make such a rigorous trade-off; consequently, they tend to result in a flatter distribution of weights. If the distribution of "true" weights is peaked (assuming that all participants know the task equally well) the techniques that tend to spread out individuals' weights should be more accurate than techniques that do not. Conversely, if the distribution of "true" weights is flat, techniques that do not tend to spread out individuals' weights should be more accurate than those that do. In summary, it was hypothesized that the relative effectiveness of different weighting techniques would depend on a technique by distribution interaction.

3.2 Method

Experiment 2 evaluated the relative effectiveness of five different techniques for obtaining an individual's weights on attributes in four tasks in the ATTACK MPS. The five weighting techniques were (1) Edward's (1977) ratio estimation technique for riskless choice (called SMART for Simple Multi-Attribute Rating Technique); (2) a paired comparison technique that forces participants to explicitly make the comparisons implicit in Edward's technique; (3) Hammond et al.'s (1975) statistical decomposition technique (uses Multiple Regression Analysis), called "policy-capturing"; (4) Keeney and Raiffa's (1976) lottery technique for risky choice; and (5) a simple, subjective rating technique, that of dividing up 100 points among the attributes. (Appendix B contains the instructions for each weighting technique for the flat five-attribute task.)

The five techniques cited above were selected for two reasons. First, as Hammond, McClelland, and Mumpower (1980) pointed out, they represent distinctly different theoretical perspectives on how an individual's weights should be obtained for multiple attributes. Consequently, their evaluation has significant theoretical and practical value, for the techniques tend to be uniquely used by different decision researchers and practitioners. Second, the techniques vary on the extent to which they force individuals to rigorously compare the worth of one attribute relative to another. Consequently, they provide the means for testing the hypothesized technique by distribution interaction.

Weights were obtained for the following four tasks in the ATTACK MPS: Command Displacement (CD), Preliminary Operations (PO), Maneuver Forward (MF), and Consolidation (C). CD and PO have five attributes; MF and C have nine attributes. In MCCRES, CD and MF have a pea:ed distribution of weights. The ratios of the most important to the least important attribute are 10.75 and 12.00, respectively, while the standard deviation of the weights is 13.79 and 8.97, respectively, on a 100-point scale. In contrast, PO and C have a relatively flat distribution of weights; the ratios of the most important to the least important attribute are 3.63 and 3.60, respectively, while the standard deviation of the weights is 7.82 and 4.89, respectively.

Systematic variation in the formal characteristics of the weighting task is a major advance over previous research in this area. Different weighting techniques have not been evaluated using an external criterion over variation in either the number of attributes or the peakness of the true weights. The research of Fischer (1977, 1979) and Von Winterfeldt and Edwards (1975) suggests that all weighting techniques will be more effective with a smaller number of attributes because of lower information processing requirements.

It was hypothesized that the relative effectiveness of the weighting techniques would depend on the actual distribution of task weights. Specifically, it was hypothesized that, regardless of task weights, SMART and paired comparisons would generate similar weights, which would have peaked distributions because the ratio, paired comparisons within both techniques, force individuals to compare rigorously the worth of one attribute relative to another, thereby spreading their weights. This should result in more accurate weights when the true distribution of weights is peaked than when it is flat.

In contrast, there is no rigorous ratio, paired comparison process for the lottery technique, policy-capturing, or dividing up 100 points. Consequently, it was hypothesized 'hat these techniques would result in a much flatter distribution of weights than generated by SMART and paired comparison techniques, regardless of actual task weights. As a result, the lottery technique, policy-capturing, and dividing up 100 points should lead to more accurate weights when the true distribution of weights is flat than when it is peaked.

In summary, the following results were predicted:

1. A main effect for the number of attributes: the smaller the number of attributes, the better the performance.

2. A technique by distribution interaction regarding the relative effectiveness of different weighting techniques. SMART and paired comparisons would result in greater accuracy when the distribution of actual weights was peaked than when it was flat; the lottery technique, policy capturing, and dividing up 100 points would be more accurate for the flat not peaked distributions of weights.

Because of the hypothesized technique by distribution interaction, there is no basis for hypothesizing a main effect for weighting techniques or for the distribution factor.

Experiment 2 was conducted during the afternoon of the first session. This was done so that the results of Experiment 2 could be used to help select the weighting techniques for Experiment 3. For example, if there were a main effect for weighting techniques, the two best techniques would have been selected for Experiment 3. As a result of the time constraints, a one-between-subjects and two-within-subjects factorial design (see Weiner, 1971, pp. 712-717) was used to test the above-stated hypotheses. In particular, all participating second lieutenants used all five weighting techniques for two tasks. Two of the groups had CD and MF; both tasks have peaked distributions, but CD has five attributes and MF has nine attributes. The other two groups had PO and C; these tasks have flat distributions, but PO has five attributes and C has nine attributes. The presentation of weighting techniques was counterbalanced across the four tasks to control for order effects.

Since each participant used all five weighting techniques for each task, it was possible to implement a sixth technique based on the other five. Specifically, each participant was shown the relative weights generated by each technique for each task. The weights for each technique were normalized to sum to 100, thereby making the weights comparable. After examining the weights for each technique, each participant specified a final set of weights. The sixth weighting technique, therefore, was a synthesis of the results of the other five.

There were three dependent variables in Experiment 2. The first one was the standard deviation of the weights generated by each technique for each task. It was hypothesized that different weighting techniques would affect the extent to which an individual's weights were spread-out. The relative effectiveness of these techniques would, in turn, depend on the spread of the actual weights for the task. Thus, the standard deviation in an individual's weights was hypothesized to be a mediating variable of relative effectiveness.

The second dependent variable was the match between the weights generated for an individual by each of the weighting techniques and the actual weights for the task. This "match" was operationalized through Pearson product-moment correlation coefficients. Specifically, the weights generated for an individual by each of the six weighting techniques were applied to the profiles used in implementing the policy-capturing technique for each task. In the terms of a linear equation, the profiles contain the values of the independent variables. For each profile, the weights were applied to the values of the independent variables to obtain a value on the dependent variable: combat readiness for that task. In this way, a set of predicted values was generated for each technique for each person; similarly, a set of "true" values was generated based on the actual weights. The values for each technique are correlated with the "true" values for the task to generate the major dependent variable, the relative effectiveness of each technique in generating an individual's MAVA weights. The result is the knowledge component (G) in the lens model equation (see Hammond et al., 1975).

There are other approaches for operationalizing the match between an individual's and the task's weights. One approach (John and Edwards, 1978b) is to use actual cases, not hypothetical profiles, and to correlate the predictions generated by the weights for each weighting technique with the actual scores for those cases, not the scores predicted by the task weights. This is the achievement measure (r_a) in the lens model equation. This approach could not be implemented in the present study, however, because performance data were generally not available for individual USMC battalions. Discussions with USMC personnel suggested, however, that this approach would have been inappropriate even if data were available because of the likelihood of large positive correlations between the attributes. Since differential weights do not result in significantly different predictions when the attributes have a high positive correlation, it would not have been possible to determine the relative effectiveness of different weighting techniques in generating the "true" task weights, the external criterion in Experiment 2.

A second approach to operationalizing the relative effectiveness of different weighting techniques is to determine if the weights generated by different techniques result in different rank-orders for non-dominated alternatives. Since there were no alternatives in the present study, this approach could not be used to operationalize effectiveness. Even if there were alternatives, this approach would have been inappropriate for the present problem. As Stillwell, Seaver, and Edwards (1981, p. 63) have noted, this approach is only appropriate for "decision problems" where the goal is to select the best alternative. It is inappropriate for "prediction problems," as in the present study, where the goal is to accurately evaluate all alternatives.

A third approach for operationalizing relative effectiveness would have been to correlate the weights directly, either in terms of a product moment correlation or a rank order correlation. The former measure is inferior to the one used in the study because the small number of attributes being correlated would result in an unacceptably large confidence level around the correlation. The latter measure is inappropriate for the present study because rank order correlations do not reflect the spread in weights, which is the hypothesized mediating variable of relative effectiveness in the present study.

The third dependent variable was the set of subjective measures analyzing how confident the participants were in the quality of their judgments in using each technique, and about how easy each technique was to use. These subjective measures are important because MAVA is routinely used when no external criterion exists, often under severe time constraints. Consequently, confidence and difficulty are important criteria for selecting weighting techniques outside laboratory conditions. It was hypothesized that participants would have the most confidence in SMART because (1) participants generate directly observable relative weights, in contrast to calculated weights in the paired comparison and policy-capturing techniques; (2) SMART appears more rigorous than dividing up 100 points; and (3) its weights are easier to understand than the probability counterparts generated by the lottery technique. It was hypothesized that dividing up 100 points would be the easiest technique, and the lottery the most difficult one.

Confidence and difficulty ratings were made on separate seven-point Likert scales. A value of one meant that participants were not confident in the quality of their judgments and found the task very difficult; a value of seven meant that they were very confident in the quality of the judgments and found the task very easy.

3.3 Results

The standard deviation of each individual's weights for each technique were inputs to a 2 (number of attributes) by 2 (types of distribution) by 6 (weighting techniques) analysis of variance (ANOVA). The type of distribution was a between subjects factor, since each participant was given either both peaked tasks or both flat tasks. The number of attributes and the weighting techniques were within subjects factors.

Table 3-1 shows the ANOVA results for the standard deviations. As hypothesized, there was a significant main effect for weighting technique. The mean standard deviations for the weights generated by SMART ($\bar{X}$ = 11.82) and paired comparisons ($\bar{X}$ = 13.11) were larger than those for policy-capturing ($\bar{X}$ = 7.74), the lottery technique (8.88), dividing-up 100 points ($\bar{X}$ = 9.0), and the final weights ($\bar{X}$ = 9.23). There was no technique by distribution interaction; the standard deviation of the weights generated by the different techniques was not dependent on the peakness of the actual distribution of weights.

There was also a main effect for the number of attributes, with the mean standard deviation for five attributes ($\bar{X}$ = 11.41) being significantly larger than for nine attributes ($\bar{X}$ = 8.51). In addition, there was a significant number by distribution interaction. For five attributes, the mean standard deviation was greater for a peaked distribution ($\bar{x}_p$ = 12.85, $\bar{x}_f$ = 9.98); for nine attributes, it was larger for a flat distribution ($\bar{x}_p$ = 7.78, $\bar{x}_f$ = 9.25). Examination of these mean scores suggests that the standard deviation of the weights for the peaked distribution was dependent on the number of attributes, while that for the flat distribution was not.

The second dependent variable was the accuracy of the different weighting techniques. The Pearson product-moment coefficients correlating the predicted scores of each individual's weighting technique with the true scores for the appropriate task were converted to Fischer Z scores. The r_z scores were inputs to a 2 (number of attributes) by 2 (types of distribution) by 6 (weighting techniques) analysis of variance (ANOVA). Again, the type of distribution was the between subjects factor; the number of attributes and the weighting techniques were within subject factors.

Notation

A = Number of Attributes
B = Type of Distribution
C = Weighting Technique

Source of Variation	df	SS	MS	F
Between subjects				
B	1	28.97	28.97	.60
Subjects within groups	18	869.86	48.33	
Within subjects				
A	1	505.68	505.68	12.17**
AB	1	281.95	281.95	6.79*
A x subjects within groups	18	747.68	41.54	
C	5	837.86	167.57	22.34***
BC	5	70.27	14.05	1.87
C x subjects within groups	90	674.64	7.50	
AC	5	49.37	9.87	.73
ABC	5	48.59	9.72	.72
AC x subjects within groups	90	1209.06	13.43	

* = $p < .05$
** = $p < .01$
*** = $p < .001$

Table 3-1

EXPERIMENT 2: ANOVA RESULTS FOR STANDARD DEVIATIONS

The results for the accuracy dependent variable strongly support the study's hypotheses (Table 3-2). The only significant results were a main effect for the number of attributes, and a weighting by distribution interaction. There were no main effects for weighting techniques or for the type of distribution of weights.

Figure 3-1 provides a pictorial representation of the main effect for the number of attributes. The performance for all weighting techniques was significantly higher for five attributes ($\bar{z}_r = 1.64$) than for nine attributes ($\bar{z}_r = 1.09$). This finding is consistent with a human information processing explanation; individuals can use all weighting techniques more effectively the smaller the number of attributes under consideration.

Figure 3-2 is a pictorial representation of the technique by distribution interaction. As hypothesized, SMART (S) and paired comparison (PC) were more effective for peaked than flat distributions of task weights. As hypothesized, the lottery technique and policy-capturing were more effective for flat than peaked distributions. The hypothesis for dividing up 100 points was, however, not confirmed; dividing up 100 points was more effective for peaked, not flat distributions.

The final set of weights, which was the sixth weighting technique, was more effective for peaked, not flat distributions. Figure 3-2 suggests that, when generating their final weights, participants relied on the weights generated by SMART, paired comparisons, and dividing up 100 points, not on those generated by policy-capturing or the lottery technique. This is consistent with the hypothesis that participants would have the most

Notation

A = Number of Attributes

B = Type of Distribution

C = Weighting Technique

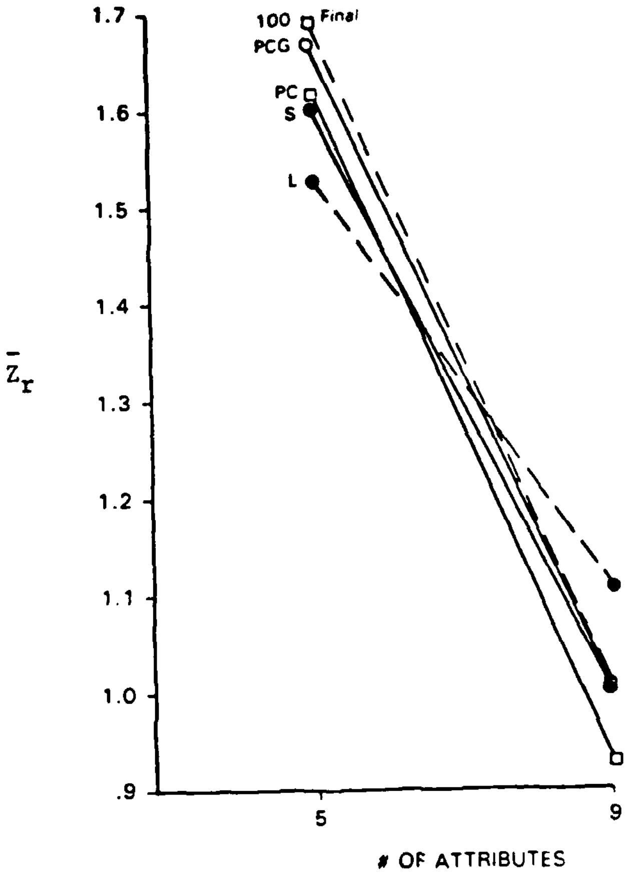
Source of Variation	df	SS	MS	F
Between subjects				
B	1	.07	.07	.08
Subjects within groups	18	14.34	.80	
Within subjects				
A	1	19.59	19.59	23.99**
AB	1	1.32	1.32	1.62
A x subjects within groups	18	14.70	.82	
C	5	.53	.11	1.21
BC	5	1.86	.37	4.24*
C x subjects within groups	90	7.89	.09	
AC	5	.54	.11	1.42
ABC	5	.60	.12	1.56
AC x subjects within groups	90	6.89	.08	

* = p < .05

** = p < .001

Table 3-2

EXPERIMENT 2: ANOVA RESULTS FOR ACCURACY

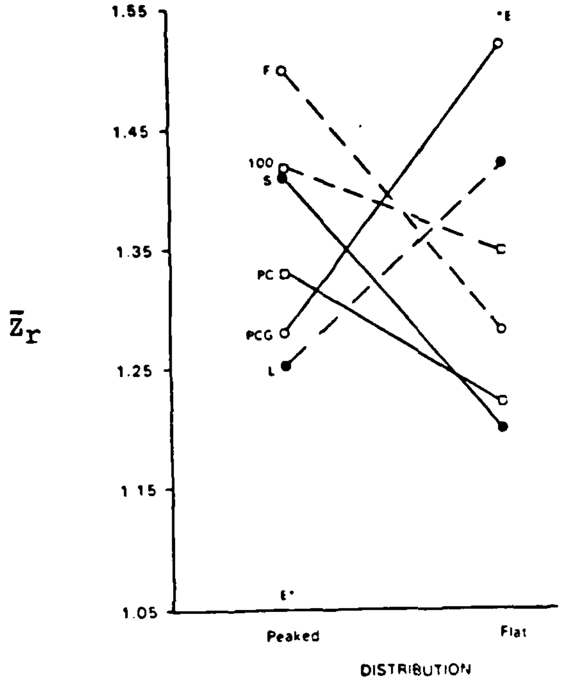


S = SMART
PC = PAIRED COMPARISON
PCG = POLICY CAPTURINC
L = LOTTERY

100 = DIVIDE UP 100 POINTS
F = FINAL
E = EQUAL WEIGHTS

Figure 3-1

EXPERIMENT 2: PICTORIAL REPRESENTATION OF THE MAIN EFFECT FOR THE NUMBER OF ATTRIBUTES



S = SMART
PC = PAIRED COMPARISON
PCG = POLICY CAPTURING
L = LOTTERY

100 = DIVIDE UP 100 POINTS
F = FINAL
E = EQUAL WEIGHTS

Figure 3-2

EXPERIMENT 2: PICTORIAL REPRESENTATION OF THE TECHNIQUE BY DISTRIBUTION INTERACTION confidence in the SMART technique and find dividing up 100 points easiest to do.

The participants' confidence ratings were inputs to a 2 (number of attributes) by 2 (types of distribution) by 5 (weighting techniques) ANOVA. The only significant effect was a main effect for weighting technique, $F(4,72) = 7.62$, $MSe = 2.87$, $p < .001$. Participants' had the most confidence in Dividing-Up 100 Points ($\bar{x} = 5.60$); next was SMART ($\bar{x} = 5.43$); then policy-capturing ($\bar{x} = 5.15$), followed by paired comparisons ($\bar{x} = 4.64$); and finally, the lottery technique ($\bar{x} = 3.76$).

The participants' difficulty ratings also were inputs to a 2 by 2 by 5 ANOVA. There were two significant effects. There was a main effect for the number of attributes $F(1,18) = 7.37$, $MSe = 1.80$, $p < .025$, with the five attribute tasks being easier ($\bar{x} = 5.17$) then the nine attribute tasks ($\bar{x} = 4.66$). There was a main effect for weighting techniques $F(4,72) = 6.59$, $MSe = 3.37$, $p < .001$, with dividing-up 100 points being easiest ($\bar{x} = 5.83$); then SMART ($\bar{x} = 5.39$); next was paired comparisons ($\bar{x} = 4.78$); followed by policy-capturing ($\bar{x} = 4.71$); and finally, the lottery technique being the most difficult ($\bar{x} = 3.86$).

3.4 Discussion

The hypotheses guiding this study were confirmed. First, the number of attributes in the task affected performance. All weighting techniques were more effective the smaller the number of attributes in the task. Second, the relative effectiveness of different techniques for obtaining an individual's MAVA weights depended on the distribution of actual task weights. Furthermore, there was a direct relationship between the ANOVA results for the standard deviation of the weights and their subsequent relative effectiveness. SMART and paired comparisons led to a greater standard deviation in the weights than did the other techniques; as hypothesized, these techniques were more effective when the task had a peaked, not a flat distribution of true weights. In contrast, policy-capturing and the lottery technique led to relative small standard deviations in the weights; both techniques were more effective when the distribution of task weights was flat, not peaked.

A perfect relationship did not exist between the standard deviation of the weights and their relative effectiveness. In this regard, two points are worth noting. First, as hypothesized, dividing-up 100 points had a mean standard deviation very close to that for policy-capturing. Yet, dividing-up 100 points was more accurate when the task weights were peaked, not flat. Second, the mean standard deviation of the correct weights for the two peaked tasks was 11.38; it was 6.36 for the two flat tasks. Although the policy capturing and lottery techniques did have the lowest mean standard deviations and, in turn, the most accurate weights for the flat tasks, the relationship did not hold for the peaked tasks. SMART had a mean standard deviation of 11.82, which was closer to the correct deviation for the two peaked tasks than any other technique. Yet, what we have referred to as the "final weights" technique was more accurate ($\bar{z}_r = 1.50$) than SMART ($\bar{z}_r = 1.42$) for the two peaked tasks. In summary, having the correct degree of spread in the weights does not imply that the weights will be assigned to the correct attributes. But in cases like the present one, where the participants can be assumed to have a good idea of the correct weights, there is definitely a relationship between the standard deviation of the weights and their accuracy, depending on the distribution of the task's actual weights.

It is important to note that all six weighting techniques led to essentially equivalent, high levels of accuracy across the four tasks. Unless one knows whether the true distribution of weights is peaked or flat, a situation which appears extremely unlikely outside the laboratory, the presumed accuracy of different weighting techniques does not seem to be a viable criterion for selecting a technique. In fact, for the four tasks used in this study, the weighting techniques were not significantly better than equal weights ($\bar{z}_r = 1.30$). As can be seen in Figure 3-2, arbitrary selection of equal weights would have resulted in the best performance for the two flat tasks and the worst performance for the two peaked tasks. These results suggest that other criteria, such as the participants' confidence in the weights obtained with the weighting technique, their confidence in the process used to generate them, the ease of using the technique, or the analyst's familiarity and confidence in the technique may be better criteria than the technique's presumed accuracy when the true distribution of weights is not known. From this perspective, SMART and dividing-up 100 points were the two best weighting techniques in the study; participants had the most confidence in the quality of the judgments generated by them, and found them easiest to use. Interestingly enough, participants had the most difficulty and the least confidence in the judgments generated by the lottery technique; yet, it was the most accurate weighting technique when the distribution of actual weights was flat.

The present study indicates that the relative effectiveness of different techniques for obtaining an individual's weights on multiple attributes depends on task characteristics. Future research studying the effectiveness of different weighting techniques must, therefore, pay serious attention to the substantive and formal characteristics of the task. The results of the present study, for example, may be limited to the content area of the four tasks and their formal characteristics, such as binary attributes, no correlations between the attributes, the number of attributes, and the distribution of actual task weights. Future research should systematically evaluate the generality of the technique by distribution interaction reported herein for different substantive and formal task characteristics, such as attributes with ordinal scales, negative correlations between the attributes, elements of risky decision making, and a broader range of both the number of attributes and the distribution of task weights.

4.0 EXPERIMENT 3: OBTAINING A GROUP'S WEIGHTS FOR MULTIPLE ATTRIBUTES

4.1 Introduction

Increasingly, MAVA is being used with groups of decision makers. This requires that MAVA weighting techniques be used in conjunction with some form of discussion technique to obtain the group's weights on the attributes. On the basis of Experiment 2, which demonstrated that different techniques affect the standard deviation of an individual's weights and, in turn, their accuracy (depending on the distribution of true weights), it was hypothesized that different discussion techniques would also affect the accuracy of group weights because they should affect their standard deviation. In particular, it was hypothesized that discussion techniques that emphasized consensus formation would result in a smaller standard deviation of group weights than discussion techniques that emphasized the weights of a few group members. If true, it would be expected that discussion techniques that emphasized consensus formation would result in more accurate group weights when the true distribution of weights is flat; and discussion techniques that focused primarily on the position of a few group members would result in more accurate group weights when the true distribution of weights is peaked.

There has been no group research on the relative effectiveness of different weighting techniques, in conjunction with different discussion techniques, over systematic variation in formal task properties. In fact, there has been very little group research studying the relative effectiveness of different weighting techniques when used in conjunction with different discussion techniques. The research conducted by Eils and John (1980) and by Rohrbaugh (1979, 1981) suggests, however, that weighting techniques do not differ in their relative effectiveness on the basis of group discussion techniques.

Eils and John (1980) evaluated the accuracy of group judgment on the basis of two factors: (1) whether or not a decomposition procedure was used; and (2) whether or not a formal group communication strategy was used. SMART was the decomposition procedure. The formal communication strategy was that proposed by Hall and Watson (1971); it attempts to break the strain toward convergence and instead, requires resolution of conflict by consensus. Eils and John (1980, p. 283) found that the "... use of the SMART decision technology significantly improved the quality of collective decisions, as did, to a lesser extent, the communication strategy." The relative effectiveness of SMART, however, was not significantly dependent on the communication strategy for two of the four criterion measures.

Rohrbaugh (1979, 1981) evaluated the relative effectiveness of the group weights and functions generated through policy-capturing procedures, depending on the type of group discussion technique. Two discussion techniques were used in each study. Social judgment analysis (e.g., see Hammond, et al., 1975) and the Delphi Technique (e.g., see Linstone and Turoff, 1975) were evaluated in the first study; social judgment analysis and the Nominal Group Technique (e.g., see Delbecq, Van de Van, and Gustafson, 1975) were evaluated in the second study.

In the social judgment analysis condition, participants were shown a pictorial representation of their weights and functions and were permitted to interact in any manner they wished until reaching a consensus position. This condition appears to be similar to the formal decomposition and formal communication strategy conditions in the study by Eils and Johns (1980). In the Delphi condition, participants did not interact directly. Rather, they revised their estimates of the weights and functions on the basis of feedback regarding statistical parameters, such as the range and median values for the weights. The Delphi group's final weights and functions were the median values of the members' individual estimates. The Nominal Group Technique (NGT) condition in the second experiment was similar to the social judgment analysis condition, except that in the NGT condition (1) group members described the reasons for their weights and functions in a round-robin fashion prior to any group discussion; and (2) the group's weights and functions were obtained through a secret balloting procedure that continued until consensus was reached. Both studies found no differences in the quality of the group decisions.

The studies by Eils and John (1980) and by Rohrbaugh (1979, 1980) suggest that the relative accuracy of the group weights, generated by different weighting techniques (SMART and policy-capturing, respectively), is not affected by the type of discussion technique. This conclusion is considered premature for a number of reasons. First, all three studies used actual data sets to generate the predicted values that were correlated with the true scores. These data sets had positive correlations between the attributes and relatively low overall predictability in terms of their multiple correlation coefficients. Such characteristics make it difficult to determine the relative effectiveness of different weighting and discussion techniques in generating the true weights, which is the external criterion in Experiment 3.

Second, none of the discussion techniques was conducted under the direction of a group leader. Instead, the techniques just varied the type of interaction between group members. Yet, when working with actual decision-making groups, MAVA procedures are always implemented under the direction of trained analysts. Analysts, like any group leader, can affect the group's weights by unintentionally controlling the extent to which different group members participate in the group discussion. Furthermore, they can readily implement group discussion techniques that either emphasize or de-emphasize the relative value of the attributes. For example, analysts at DDI often use weighting and discussion techniques that emphasize ratio-paired comparisons between attributes. On the basis of the results of Experiment 2, the relative accuracy of the group weights generated by such techniques should depend on the formal characteristics of the task.

Third, all the discussion techniques used by Eils and John (1980) and by Rohrbaugh (1979, 1981) were oriented toward consensus formation. In all cases, participants received feedback that would permit them to compare their weights (and functions) with those of other group members. Such comparison increases the probability that each group member will have an equal say in the group's weights, thereby facilitating a consensus position that is based on compromising extreme differences in weights. Such a process should result in a flatter distribution of weights than generated by discussion procedures which do not permit group members to compare their weights. When using MAUA with actual decision-making groups, time constraints often force analysts to obtain group weights without having individual group members first specify their own position. Under such conditions, one or two group members can control the discussion and have their weights adopted by the group with minimal, if any, compromise by other group members. The result might be a more peaked distribution of weights than that generated by any of the discussion procedures used by Eils and Johns (1980) and by Rohrbaugh (1979, 1981), all of which emphasize compromise and consensus formation.

Assuming that all group members are equally likely to know the correct weights, the following hypotheses were proposed on the basis of the results for Experiment 2. First, the relative effectiveness of different discussion techniques used in obtaining group weights will depend upon the peakness of the true distribution of weights. Discussion techniques that emphasize consensus formation will be most effective when the true distribution of weights is peaked, not flat. Second, this effect will be compounded further by the type of weighting technique used by the analyst to obtain group weights. Weighting techniques, like SMART, which emphasize ratio-paired comparisons between attributes, will be most effective with discussion techniques that do not emphasize consensus formation when the true distribution weights is peaked. In contrast, weighting techniques, like policy-capturing, will be most effective with discussion techniques that emphasize consensus formation when the true distribution of weights is flat. All of these hypotheses are based on the assumption that the standard deviation of the group weights is a viable mediating variable of relative effectiveness.

4.2 Method

Experiment 3 evaluated the relative effectiveness of two weighting techniques (in conjunction with two discussion techniques) for obtaining group weights for the same four tasks used in Experiment 2. The two weighting techniques were SMART and policy capturing. The two discussion techniques were the

Nominal Group Technique (NGT), which emphasizes consensus formation, and a leader-directed technique that does not emphasize consensus formation. The four tasks varied two formal characteristics: the number of attributes and the peakness of the distribution of true weights. Since the weighting techniques and tasks are described in Experiment 2, only the discussion techniques are described here. (Appendix C contains the analysts' instructions for each weighting technique.)

In brief, the NGT condition emphasized three aspects that differed from the leader-directed condition. First, participants in the NGT were given an opportunity to tell group members their weights (which are written on the board), and the reasons for them, in a round-robin manner that did not permit discussion until all group members had finished their presentation. In contrast, participants in the leader-directed technique were not given an opportunity to tell group members their weights; their data sheets for the weighting technique were collected before the beginning of the group's discussion and the analysts were told not to focus on individual members weights. Second, the NGT emphasized discussion among group members and not with the analyst; the analyst's primary role was to help group members implement consensus formation guidelines. In contrast, analysts took an extremely active role in focusing the group's discussion in the leader-directed technique; the discussion was primarily between group members and the analyst, and not among group members. Third, the group weights in the NGT were determined arithmetically by taking the geometric means of a second set of weights that were generated individually by each group member after the group discussion. In contrast, group weights in the leader-directed technique were obtained during the discussion.

There were two dependent variables in Experiment 3. The first one was the standard deviation of the weights generated by each combination of weighting and discussion techniques for each of the four tasks. It was hypothesized that different techniques would affect the extent to which group weights were spread-out. In particular, it was hypothesized that the leader-directed technique would result in group weights with a larger standard deviation than those generated by the Nominal Group Technique. A further hypothesis was that SMART would result in group weights with a larger standard deviation than those generated by policy capturing. The relative effectiveness of these techniques would, in turn, depend on the spread of the actual weights for the task. Thus, the standard deviation of a group's weights was hypothesized to be a mediating variable of relative effectiveness, the second dependent variable in the study.

Two other hypotheses were made about the standard deviations of the group weights. First, it was hypothesized that the mean standard deviation would be larger for tasks with five rather than nine attributes. This hypothesis was based on the results obtained in Experiment 2, where this main effect was found for individual participants' weights. Second, it was hypothesized that the standard deviation of the group weights would be larger for tasks with peaked rather than flat distributions. Although this hypothesis was not supported in Experiment 2, it still appears appropriate if one assumes that the participants have a good idea of the relative importance of the attributes for all four tasks and, therefore, generally know the spread of the weights.

The relative effectiveness of the weights generated by different combinations of weighting and discussion techniques was measured by the match between the group weights generated by each weighting-discussion technique combination and the true weights for the task. As in Experiment 2, this "match" was operationalized through Pearson product-moment correlations. In this experiment, the correlations were between the predicted, combat readiness values (for the task profiles) generated by the group weights obtained for each weighting-discussion technique combination, and the true "combat readiness" values generated by the actual weights for the task.

The following four hypotheses were proposed regarding the relative effectiveness of different weighting and discussion techniques:

1. On the basis of the results for Experiment 2, it was hypothesized that both weighting techniques and both discussion techniques would be more effective for tasks with five attributes than tasks with nine attributes.

2. A weighting technique by distribution interaction - SMART will result in more accurate group weights for peaked, not flat distributions; policy-capturing will result in more accurate group weights for flat, not peaked distributions.

3. A discussion technique by distribution interaction - NGT will result in more accurate weights for flat, not peaked distributions; the leader-directed techniques will result in more accurate weights for peaked, not flat distributions.

4. A weighting by discussion by distribution interaction - SMART will result in the most accurate group weights when used with the leader-directed discussion technique for the peaked distribution; it will result in the least accurate group weights when used with the leader-directed discussion technique for the flat distribution. In contrast, it was

predicted that policy capturing would result in the most accurate group weights with the NGT discussion technique for the flat distributions, and the least accurate group weights with NGT for the peaked distribution.

Experiment 3 was implemented during the second half of the last five sessions with the participating USMC second lieutenants. The participants remained in the four, five-person groups to which they had been assigned for the structuring study (Experiment 1) that morning. Each group received four of the sixteen conditions of the experiment: 2 (discussion techniques) by 2 (weighting techniques) by 2 (number of attributes) by 2 (type of distribution). The four conditions for each group involved a single discussion technique, either the NGT or the leader-directed technique, to obtain the group weights for all four tasks. Each group used SMART for two tasks and policy-capturing for two tasks. For three replications of the experiment, groups were assigned to blocks so that the interaction between distribution and weighting technique was confounded with block differences. For the remaining two replications, blocks were constructed which confounded the three-way interaction between number, distribution, and weighting technique. Between- and within-group tests were combined for those effects, with two estimates of size.

Prior to implementing the group discussion technique, weights were obtained for each group member. This permitted two additional analyses. First, it was possible to perform a replication study of Experiment 2 for SMART and policy capturing, but now the sample size was increased; almost all groups had five members. On the basis of the results for Experiment 2, it was hypothesized that (1) both weighting techniques would result in more accurate weights for the five attribute, than the nine attribute tasks; and (2) that SMART would be more effective than policy capturing for peaked distributions; and (3) that policy capturing would be more effective than SMART for flat distributions.

Second, two analyses of variance were performed on the residuals of the correlation between the standard deviations, and the accuracy ($\bar{z}_r$) of the group weights before and after discussion, respectively. These analyses permitted DDI to supplement the analysis of variance performed on the post discussion group weights by evaluating how the discussion techniques affected the group weights. No analysis was performed on the prediscussion group weights (determined arithmetically by taking the geometric mean of the participants' weights for the attributes, and normalizing them to sum to 100) because that analysis duplicated the analysis performed for individual participants. The only difference procedurally would have been that individuals were randomly clustered into groups of five, thereby greatly reducing the sample size for the analysis.

4.3 Results

The first analysis presented herein is for the standard deviations and performance measures for the individuals' weights. Specifically, the standard deviation of each individual's weights, and the z scores for the correlations between the predicted values of each individual's weights with the true scores for the appropriate task, were inputs to separate 2 (number of attributes) by 2 (types of distribution) by 2 (weighting techniques) analysis of variance (ANOVA) designs. Since participants were in different groups with different analysts leading them, the variance due to group effects was partitioned into a between-group factor. The number of attributes, the peakness of the distribution, and the weighting technique were all within-group factors. Combined tests were perfo~med for those effects whose variance was determined by both between-group and within-group factors.

Table 4-1 presents the ANOVA results for the standard deviations of the individuals' weights. There were a number of statistically significant effects, but the main effects for the weighting technique, and the number of attributes accounted for most of the variance. In particular, the mean standard deviation for SMART ($\bar{x}$ = 12.42) was significantly larger than that for policy capturing ($\bar{x}$ = 7.08); and the mean standard deviation for the five attribute tasks ($\bar{x}$ = 11.30) was significantly larger than that for the nine attribute tasks ($\bar{x}$ = 8.20), as obtained in Experiment 2. In addition, the mean standard deviation for tasks with peaked distributions ($\bar{x}$ = 10.44) was significantly larger than that for tasks with flat distributions ($\bar{x}$ = 9.06), as hypothesized. The significant interaction between the weighting technique and the number of attributes, which was not hypothesized, indicates that the difference in the standard deviations between SMART and policy capturing was larger for the five attributes than for the nine attribute tasks. The significant interaction between the peakness of the distribution and the number of attributes, which also was obtained in Experiment 2, reflects the fact that the difference in the standard deviations between the peaked and flat tasks was larger for the five than for the nine attribute tasks. Although statistically significant, both interactions accounted for only 1.5% of the variance in the ANOVA.

Table 4-2 presents the ANOVA results for accuracy (z_r). These results replicate those found in Experiment 2. There was a main effect for the number of attributes; participants' weights were more accurate for the five attribute tasks ($\bar{z}_r$ = 1.47) than the nine attribute task ($\bar{z}_r$ = 1.04). And there was

SOURCE	df	SS	MS	F
Between Subjects	95	2159.321		
Groups	3	165.921		
Replications	1	21.04945		
BC	1	71.17349	71.17349	
ABC	1	73.69806	73.69806	
Subjects with Group	92	1993.4	21.66739	
Within Subjects	288	7548.385		
Number (A)	1	927.4645	927.4645	73.09***
Distribution (B)	1	181.6926	181.6926	14.32***
Weighting Tech. (C)	1	2740.5957	2740.5957	215.99***
AB	1	55.815	55.815	4.40*
AC	1	65.175	65.175	5.14**
BC	1	2.79727	2.79727	
ABC	1	9.36821	9.36821	
Residual	281	3565.4768	12.6885	

COMBINED TESTS OF BC AND ABC

SOURCE	df	MS	F
BC	1	28.0503	1.75
ABC	1	33.1268	2.07
Error	373	16.0046	

* = p <.05
** = p <.025
*** = p <.001

Table 4-1

ANOVA FOR STANDARD DEVIATIONS
OF INDIVIDUAL PARTICIPANTS' WEIGHTS

SOURCE	df	SS	MS	F
Between Subjects	95	19.41965		
Groups	3	1.67813		
Replications	1	.23436		
BC	1	1.42776	1.42776	
ABC	1	.01601	.01601	
Subjects with Group	92	17.74152	.19284	
Within Subjects	288	75.96645		
Number (A)	1	17.13660	17.13660	93.31***
Distribution (B)	1	.09375	.09375	.51
Weighting Tech. (C)	1	.42667	.42667	2.32
AB	1	5.07840	5.07840	27.65***
AC	1	.89707	.89707	4.88*
BC	1	.66317	.66317	
ABC	1	.06623	.06623	
Residual	281	51.60456	.18365	

COMBINED TESTS OF BC AND ABC

SOURCE	df	MS	F
BC	1	1.03614	5.51**
ABC	1	.04173	.22
Error	373	.18813	

* = p <.05
** = p <.025
*** = p <.001

Table 4-2

ANOVA FOR ACCURACY (Z_r) OF
INDIVIDUAL PARTICIPANTS' WEIGHTS

a significant weighting technique by distribution interaction. When the distribution was peaked, SMART was more accurate ($\bar{z}_r$ = 1.28) than policy capturing ($\bar{z}_r$ = 1.21); when the distribution was flat, policy capturing ($\bar{z}_r$ = 1.38) was more accurate than SMART ($\bar{z}_r$ = 1.16).

Experiment 3 also obtained two significant findings not obtained in Experiment 2. First, there was a significant interaction between the number of attributes and the peakness of the distribution. For five attributes, performance was better for the task with the peaked ($\bar{z}_r$ = 1.57) than the flat ($\bar{z}_r$ = 1.37) distribution; for nine attributes, performance was better for the task with the flat ($\bar{z}_r$ = 1.17), than the peaked ($\bar{z}_r$ = .92) distribution. Second, there was an interaction between the type of weighting technique and the number of attributes. SMART ($\bar{z}_r$ = 1.49) was slightly more effective than policy capturing ($\bar{z}_r$ = 1.46) for the five attribute tasks, but policy capturing ($\bar{z}_r$ = 1.13) was more effective than SMART ($\bar{z}_r$ = .96) for the nine attribute tasks.

Table 4-3 presents the ANOVA results for the standard deviations of the group weights obtained after discussion. Discussion was a between-group factor because each group used either the leader-directed technique or the NGT for all tasks; the number of attributes, the peakness of the distribution, and the weighting technique were within-group factors. All four hypotheses were confirmed in the study. First, the leader-directed technique ($\bar{x}$ = 11.46) resulted in a larger mean standard deviation of the group weights than did the NGT ($\bar{x}$ = 7.83). Second, SMART ($\bar{x}$ = 11.62) resulted in a larger mean standard deviation than policy capturing ($\bar{x}$ = 7.85). Third, the mean standard deviation of the group weights for the five attribute tasks ($\bar{x}$ = 11.03) was larger than that for

SOURCE	df	SS	MS	F
Between Groups	19	684.066		
Blocks	7	481.929		
Discussion Type (A)	1	263.429	263.429	15.64***
CD (from #5,6)	1	57.072	57.072	
ACD (from #5,6)	1	49.939	49.939	
BCD (from #7,8)	1	0.188	0.188	
ABCD (from #7,8)	1	46.489	46.489	
Replications	2	64.812	32.406	
Groups w/Block	12	202.137	16.845	
Within Groups	60	1455.069		
Number (B)	1	155.264	155.264	11.34**
Distribution (C)	1	118.950	118.950	8.69*
Weighting Tech. (D)	1	313.592	313.592	22.90***
AB	1	10.232	10.232	.75
AC	1	9.119	9.119	.67
AD	1	5.666	5.666	.41
BC	1	165.514	165.514	12.09**
BD	1	0.280	0.280	.02
CD (from #7,8)	1	7.990	7.990	
ABC	1	1.313	1.313	.10
ABD	1	3.482	3.482	.25
ACD (from #7,8)	1	4.097	4.097	
BCD (from #5,6)	1	29.704	29.704	
ABCD (from #5,6)	1	0.049	0.049	
Residual	46	629.817	13.692	

COMBINED TESTS

SOURCE	df	MS	F
CD	1	29.997	1.99
ACD	1	24.651	1.63
BCD	1	16.386	1.08
ABCD	1	20.871	1.38
Error	58	15.1057	

* = p < .01
** = p < .005
*** = p < .001

Table 4-3

ANOVA FOR STANDARD DEVIATIONS OF GROUPS' WEIGHTS the nine attribute tasks ($\bar{x}$ = 8.25). Fourth, the mean standard deviation was significantly larger for tasks with peaked ($\bar{x}$ = 10.86) than flat ($\bar{x}$ = 8.42) distributions. In addition, a significant number by distribution interaction was obtained, although it was not hypothesized. When the tasks had five attributes, the mean standard deviation was larger for the task with the peaked ($\bar{x}$ = 13.69), not flat ($\bar{x}$ = 8.37) distribution; when the tasks had nine attributes, the mean standard deviation was larger for the task with the flat ($\bar{x}$ = 8.47), not peaked ($\bar{x}$ = 8.03) distribution.

Table 4-4 presents the ANOVA results for the accuracy (z_r) of the group weights after discussion. In contrast to the results obtained for the standard deviations, only one of the four hypotheses was confirmed for group accuracy. The hypothesized main effect for the number of attributes was obtained; group accuracy was significantly larger for the five ($\bar{z}_r$ = 1.71) than nine ($\bar{z}_r$ = 1.14) attribute tasks. The results did not confirm the hypothesized weighting technique by distribution, discussion technique by distribution, and weighting by discussion by distribution interactions. A significant number by distribution interaction was, however, obtained. When the tasks had five attributes, accuracy was better for the task with the peaked ($\bar{z}_r$ = 1.87) not flat ($\bar{z}_r$ = 1.56) distribution; when the task had nine attributes, accuracy was better for the task with the flat ($\bar{z}_r$ = 1.25) than the peaked ($\bar{z}_r$ = 1.02) distribution.

The group accuracy results were quite surprising for two reasons. First, the hypothesized weighting technique by distribution interaction was obtained for the individual participants' weights; second, the hypothesized results were obtained for the standard deviations of the groups' weights. An analysis of variance was performed on the residuals of the

SOURCE	df	SS	MS	F
Between Groups	19	4.6486		
Blocks	7	0.8415		
Discussion Type (A)	1	0.389	0.389	1.22
CD	1	0.014	0.014	
ACD	1	0.069	0.069	
BCD	1	0.263	0.263	
ABCD	1	0.024	0.024	
Replications	2	0.082	0.041	
Groups w/Block	12	3.8071	0.3173	
Within Groups	60	16.0557		
Number (B)	1	6.705	6.705	48.52***
Distribution (C)	1	0.030	0.030	.22
Weighting Tech (D)	1	0.202	0.202	1.46
AB	1	0.006	0.006	.04
AC	1	0.456	0.456	3.30
AD	1	0.032	0.032	.23
BC	1	1.453	1.453	10.51**
BD	1	0.290	0.290	2.10
CD	1	0.248	0.248	
ABC	1	0.048	0.048	.35
ABD	1	0.087	0.087	.63
ACD	1	0.045	0.045	
BCD	1	0.075	0.075	
ABCD	1	0.022	0.022	
Residual	46	6.3567	0.1382	

COMBINED TESTS

SOURCE	df	MS	F
CD	1	0.1770	.92
ACD	1	0.0523	.27
BCD	1	0.1320	.69
ABCD	1	0.0278	.14
Error	58	0.1925	

** = p < .005
*** = p < .001

Table 4-4

ANOVA FOR ACCURACY (Z_r) OF GROUPS' WEIGHTS correlation between the accuracy ($\bar{z}_r$) of the group weights, before and after discussion, in an effort to better understand the effects of discussion on the accuracy of the group weights. Table 4-5 presents the results of this analysis. There were two significant effects. First, there was a significant main effect for the number of attributes. Accuracy was better after discussion for the five attribute tasks (($\bar{z}_{r,a}$ = 1.71, $\bar{z}_{r,b}$ = 1.67) and worse after discussion for the nine attribute task ($\bar{z}_{r,a}$ = 1.14, $\bar{z}_{r,b}$ = 1.25).

Second, there was a significant weighting technique by distribution interaction. For tasks with peaked distributions, performance decreased with SMART ($\bar{z}_{r,a}$ = 1.35, $\bar{z}_{r,b}$ = 1.42) and increased with policy capturing ($\bar{z}_{r,a}$ = 1.54, $\bar{z}_{r,b}$ = 1.34). For tasks with flat distributions, performance increased with SMART ($\bar{z}_{r,a}$ = 1.40, $\bar{z}_{r,b}$ = 1.34) and decreased with policy capturing ($\bar{z}_{r,a}$ = 1.41, $\bar{z}_{r,b}$ = 1.67). These results would be consistent with the hypothesis that the standard deviation of the weights is a mediating variable of relative effectiveness under the following condition: the standard deviation of the group weights generated by SMART decreased after discussion, while they increased for policy capturing. Examination of the standard deviation of the group weights after discussion indicates they increased for both weighting techniques, and for both peaked and flat distributions. These results led to rejection of the hypothesis that the standard deviation of group weights, in conjunction with the distribution of actual task weights, was a mediating variable of their relative effectiveness.

SOURCE	df	SS	MS	F
Regression	1	13.050	13.050	132.99
Residual	78	7.654	.098	
Between Groups	19	2.258		
Blocks	7	0.914		
Discussion Type (A)	1	0.025	0.025	0.22
CD	1	0.676	0.676	
ACD	1	0.020	0.020	
BCD	1	0.026	0.026	
ABCD	1	0.000	0.000	
Replications	2	0.167		
Groups w/Block	12	1.344	0.112	
Within Groups	59	5.396		
Number (B)	1	0.352	0.352	4.40*
Distribution (C)	1	0.279	0.279	3.49
Weighting Tech (D)	1	0.005	0.005	0.06
AB	1	0.109	0.109	1.36
AC	1	0.024	0.024	0.30
AD	1	0.046	0.046	0.58
BC	1	0.020	0.020	0.25
BD	1	0.003	0.003	0.04
CD	1	0.591	0.591	
ABC	1	0.045	0.045	0.56
ABD	1	0.152	0.152	1.90
ACD	1	0.107	0.107	
BCD	1	0.046	0.046	
ABCD	1	0.011	0.011	
Residual	45	3.606	0.080	

COMBINED TESTS

SOURCE	df	MS	F
CD	1	0.626	6.73**
ACD	1	0.071	0.76
BCD	1	0.038	0.41
ABCD	1	0.006	0.06
Error	57	0.093	

* = $p < .05$
** = $p < .025$

Table 4-5

ANOVA FOR RESIDUALS OF ACCURACY (Z_r)

4.4 Discussion

The results of this study supported the findings reported by Eils and John (1980) and by Rohrbaugh (1979, 1981) that discussion techniques do not affect the accuracy of group weights. In addition, the present study showed that different weighting techniques also do not affect the accuracy of group weights. The only factors that did so were task characteristics. In particular, group accuracy was significantly higher for the five than for the nine attribute tasks. Furthermore, when the task had five attributes, group accuracy was higher for the task with a peaked, not flat distribution. When the task had nine attributes, group accuracy was higher for the task with a flat, not peaked distribution.

Although these results need to be confirmed using other tasks, they suggest that decision analysts should pay more attention to task characteristics than to weighting or discussion techniques. When developing multi-attributed hierarchies, decision analysts can control the number of attributes at each node by how they cluster attributes. The present study indicates fewer attributes are better. Future research with external criteria is required, however, to determine whether partitioning a set of attributes into two smaller subsets results in more accurate group weights for the larger set of attributes. Furthermore, that research might vary the way the attributes are partitioned into subsets. For example, does partitioning the set into two subsets, one with most important attributes and one with the least important attributes, result in more accurate group weights than subsets where the attributes are assigned randomly? The former approach is often used by decision analysts at DDI, but there is no empirical research demonstrating that it results in more accurate group weights than the latter approach.

There was no support for the proposition that the standard deviation of the weights, in conjunction with the distribution of actual task weights, was a mediating variable of the relative effectiveness of group weights. This result was quite surprising for two reasons. First, the standard deviations were a mediating variable of the relative effectiveness of individual participants' weights. SMART led to a much larger standard deviation of the individual weights than did policy capturing. In turn, the weights generated by SMART were more accurate than those generated by policy capturing when the distribution of task weights was peaked, but less accurate when the distribution was flat. Second, all of the hypotheses were confirmed for the standard deviations of the group weights. The leader-directed discussion technique led to a larger standard deviation of the group weights that did the NGT. SMART led to a larger standard deviation than policy capturing. Yet, the discussion by distribution, weighting by distribution, and discussion by weighting by distribution interactions were not significant for the accuracy of the group weights.

Having the correct spread in the weights does not, of course, ensure that the weights will be put on the correct attributes. Future research may even demonstrate that the observed relationship between the standard deviation of an individual's weights and their relative effectiveness does not hold for tasks with different formal and substantive properties. But Experiment 3 strongly suggests that this relationship will not hold for group weights obtained through discussion.

It is not clear how discussion was affecting the accuracy of the group weights in Experiment 3. One hypothesis is that discussion had no effect at all, and that group accuracy was determined by regression to the mean. The analysis of variance performed on the residuals between the accuracy ($\bar{z}_r$) of the group weights before and after discussion provide some support for this hypothesis, because there was a significant weighting by distribution interaction with effects in the appropriate direction. For SMART, performance decreased for tasks with peaked distributions and increased for tasks with flat distributions, while just the opposite occurred with policy capturing. Other aspects of the ANOVA performed on the residuals do not, however, support the hypothesis of regression to the mean. For example the significant main effect for the number of attributes occurred because performance for the five attribute tasks, which was high before discussion, became even higher after discussion; while performance for the nine attribute tasks, which was low before discussion, became even lower after discussion. In addition, performance for the leader-directed discussion technique grew better, not worse, for tasks with a peaked distribution, while it grew worse, not better, for tasks with a flat distribution, contrary to the hypothesis of regression to the mean.

In order to better understand the effect of discussion on group accuracy, future research should not only vary formal and substantive task properties, but vary characteristics of the broader decision problem and decision-making group. Regarding the first point, the present study showed that the leader-directed technique and the NGT were equally effective for tasks with binary attributes, high predictability, zero correlations between the attributes, and no distinct alternatives. Future research should investigate whether the present findings generalize to different substantive tasks with different formal characteristics, such as ordinal atttributes, low predictability, negative inter-attribute correlations, and distinct alternatives. This latter task property also would permit the use of different measures of effectiveness, such as the rank order of the alternatives.

Regarding the second point, that of the decision problem and group, the decision problem in the present study was to assign weights to multiple attributes; the group was a homogeneous group of peers. For a different set of decision problems and groups, Stumpf, Freedman, and Zand (1979) have shown that characteristics of the decision-making problem and group do affect the relative effectiveness of different discussion techniques. Future research could readily investigate, for example, whether the status of group members affects group accuracy, and if so, what discussion techniques minimize its negative effect. It is hypothesized that group status should have a negative effect when the distribution of task weights is peaked because mistakes have a greater detrimental effect for peaked than for flat distributions. Consistent with this position, the NGT should be more effective than the leader-directed technique at minimizing this negative effect, because it provides norms for group interaction.

5.0 USING MULTI-ATTRIBUTE VALUE ASSESSMENT TECHNIQUES AS DIAGNOSTIC TOOLS FOR TRAINING

The goal of the Infantry Officers Course (IOC) at The Basic School is to train Marine Corps second lieutenants not only in how to perform combat requirements and tasks, but also in the relative importance of and reason for performing them. Said differently, the goal is to produce thinking as well as fighting Marines. In this section, we will briefly describe how MAVA structuring techniques were used to identify what graduating IOC students thought were the important requirements and tasks in performing the ATTACK MPS in MCCRES. By doing so, Marine Corps training personnel were able to identify the relative strengths and weaknesses of their curriculum.

Experiment 1 evaluated the relative effectiveness of two MAVA structuring techniques. The techniques were implemented by trained decision analysts, each of whom led a group of five second lieutenants who had just graduated from the IOC. The participants represented a cross-section of performance levels in their graduating class, and they were assigned randomly to their groups. The second lieutenants were tasked to develop a structured evaluation system for measuring an infantry company's performance in attacking the enemy during a combat readiness evaluation. Their evaluation structure, like the ATTACK MPS, was to include those specific actions required to make the final assault in daylight. Each group's completed evaluation structure was then scored against MCCRES, the external criterion, by an independent evaluator in the Operations Division of the Plans, Policies, and Operations Department at Marine Corps Headquarters who had participated in writing the MCCRES volume for infantry units. By doing so, DDI analysts were able to identify how well the second lieutenants knew the performance requirements for each task in the ATTACK MPS in MCCRES. Since the IOC does not teach MCCRES through rote memorization, it represents a good external criterion for evaluating the knowledge of graduating second lieutenants.

Participants performed well, meeting Marine Corps performance standards for the ATTACK MPS according to assessments by the independent evaluator. On the average for all seven sessions, the groups identified 44% of the requirements in the ATTACK MPS. These requirements accounted, on an average, for 48% of the possible cumulative utility for the criterion. Performance was considerably better for the first six tasks in the ATTACK MPS, which are the tasks in this MPS for which USMC second lieutenants received more training. Participants identified, on the average, 57% of the requirements, representing 62% of the possible cumulative utility for these six tasks. Furthermore, they almost always identified the most important requirement for each of these tasks, obtaining 94% of the cumulative utility for them.

MAVA structuring techniques can be used to identify the relative strengths and weaknesses of a training program in addition to evaluating its overall effectiveness. For example, although the performance of second lieutenants participating in the first session of Experiment 1 met Marine Corps standards, they did not identify as many requirements as the evaluator and their instructor expected they would for the Consolidation task in the ATTACK MPS. The curriculum for the Consolidation portion of the IOC was changed as a result of this finding. Examination of Figure 5-1 shows that participants in the second and subsequent sessions of Experiment 1 identified considerably more requirements for the Consolidation task than did participants in the first session. It is important to emphasize that it cannot be concluded statistically that the

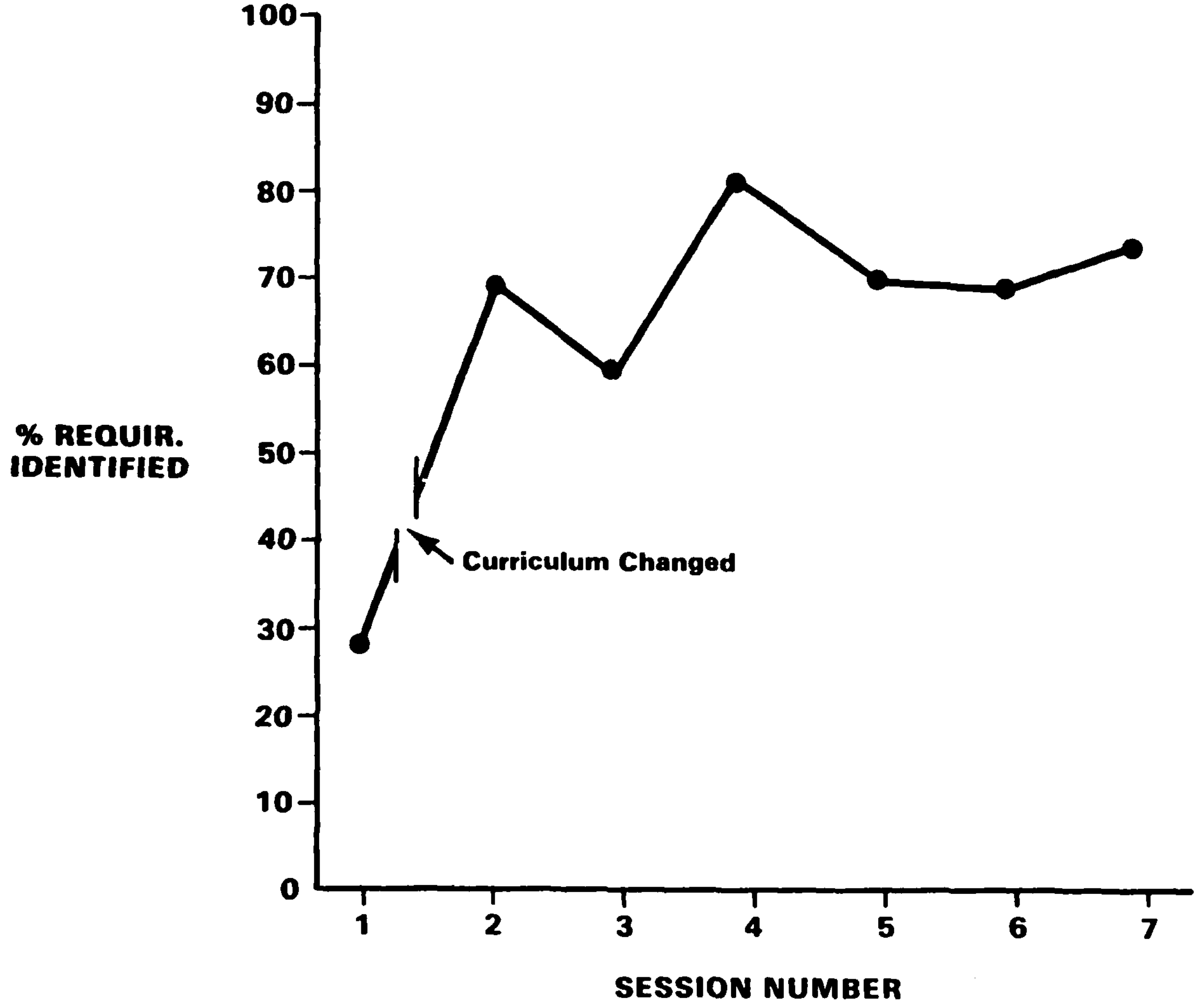


Figure 5-1

PERFORMANCE FOR THE CONSOLIDATION TASK

curriculum change was the sole or principal cause of this improved performance, since there was not a control group where second lieutenants used the old curriculum. Other factors which were not under the control of the analyst, or which may not have been evident, may be the actual cause of the observed improvement in performance. Examination of Figure 5-2, which shows the participants' performance for the Assault task (a task for which no curriculum changes were made during the course of the study) suggests, however, that the curriculum change is a viable hypothesis for explaining the improved performance for the Consolidation task. Figure 5-2 does not show the marked increase in performance after the first session that is evident in Figure 5-1.

Other curriculum changes also occurred as a result of other sessions in Experiment 1, clearly demonstrating that Marine Corps training personnel found the multi-attribute structuring techniques to be effective diagnostic tools. Weighting techniques were not used as diagnostic tools for facilitating curriculum changes in the present study. However, the results of Experiments 2 and 3 suggest that they might have been useful. For although performance for all four tasks was quite high (the mean z-score of the correlation across all four tasks was $\bar{z}_r = 1.35$ in Experiment 2, which translates into a Pearson product moment correlation of $r = .875$), performance for the two tasks with a flat distribution of relative weights (i.e., Preliminary Operations and Consolidation) would have been even better if participating second lieutenants had assigned equal weights to the requirements. Future research should systematically investigate the effectiveness of multi-attribute weighting techniques as diagnostic tools for training.

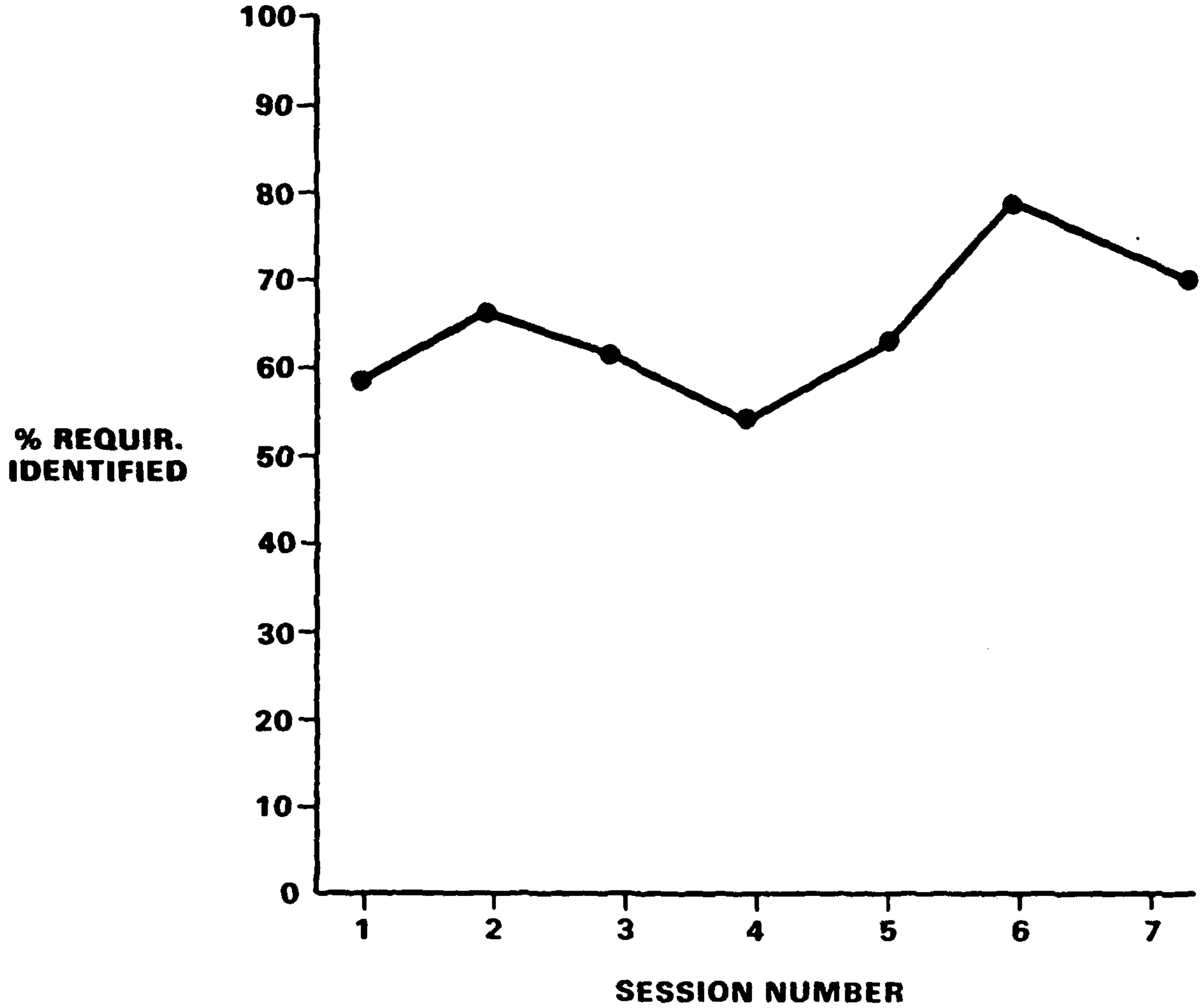


Figure 5-2

PERFORMANCE FOR ASSAULT TASK

In closing, it is important to emphasize that MAVA structuring techniques can be used as diagnostic tools for most, if not all, training programs that focus on what students think is important rather than on, or in addition to, how they perform specified tasks. Furthermore, MAVA techniques can be used in a cost-efficient fashion, with a representative group of experts in a particular field, to develop the external criterion against which to measure students' knowledge and performance, as they were used to develop MCCRES.

REFERENCES

Brunswik, E. Representation design and probabilitistic theory in a functional psychology. Psychological Review, 1955, 62, 193-217.

Cook, R. L., & Stewart, T. R. A comparison of seven methods for obtaining subjective descriptions of judgmental policy. Organizational Behavior and Human Performance, 1975, 13, 31-45.

Delbecq, A. L. , Van de Van, A. H., & Gustafson, D. H. Group techniques for program planning. Glenview, IL: Scott, Foresman, 1975.

Edwards, W. How to use multiattribute utility measurement for social decision making. IEEE Transactions on Systems, Man, and Cybernetics, 1977, SMC-7, 326-340.

Eils, L. C., III, & John, R. S. A criterion validation of multiattribute utility analysis and of group communication strategy. Organizational Behavior and Human Performance, 1980, 25, 268-288.

Einhorn, H., & Hogarth, R. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 1975, 13, 171-192.

Fischer, G. W. Convergent validation of decomposed multiattribute utility assessment procedures for risky and riskless decisions. Organizational Behavior and Human Performance, 1977, 18, 295-315.

Fischer, G. W. Utility models for multiple objective decisions: Do they accurately represent human preferences? Decision Sciences, 1979, 10, 451-479.

Fischhoff, B. Decision analysis: Clinical art or clinical science. In L. Sjoberg and J. Wise (Eds.), Proceedings of the Sixth Research Conference on Subjective Probability, Utility and Decision-Making, (Warsaw, 1977).

Fischhoff, B., Slovic, P., & Lichtenstein, S. Fault trees: Sensitivity of estimated failure probabilities to problem representation. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 330-344.

Gettys, C. F., Manning, C., & Casey, J. T. *An evaluation of human act generation performance*. Technical Report TR 15-8-81. Norman, OK: Decision Processes Laboratory, University of Oklahoma, August 1981.

Hall, J., & Watson, W. H. The effects of a normative intervention on group decision-making Performance. *Human Relations*, 1971, *23*, 299-317.

Hammond, K. R. Probabilitistic functionalism: Egon Brunswik's integration of the history, theory, and method of psychology. In K. R. Hammond (Ed.), *The Psychology of Egon Brunswik*. New York: Holt, Rinehart, and Winston, 1966.

Hammond, K. R. *Principles of organization in intuitive and analytical cognition*. Technical Report 231. Boulder, CO: Center for Research on Judgment and Policy, University of Colorado, 1981.

Hammond, K. R., McClelland, G. H., & Mumpower, J. *Human judgment and decision-making: Theories, methods, and procedures*. New York: Hemisphere/Praeger, 1980.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. Social judgment theory. In M. F. Kaplan and S. Schwartz (Eds.), *Human judgment and decision processes*. New York: Academic Press, 1975.

Hammond, K. R., & Wascoe, N.E. (Eds.). *New direction for methodology of social and behavioral science: Realizations of Brunswik's representative design*. San Francisco: Jossey-Bass, 1980.

John, R. S., Collins, L., & Edwards, W. *A comparison of importance weights for MAUA derived from holistic, indifference, direct subjective and rank order judgments*. Technical Report 80-4. Los Angeles: Social Science Research Institute, University of Southern California, 1980.

John, R. S., & Edwards, W. *Importance weight assessment for additive riskless preference functions: A review*. SSRI Research Report 78-5. Los Angeles: Social Science Research Institute, University of Southern California, 1978a.

John, R. S., & Edwards, W. *Subjective versus statistical importance weights: A criterion validation*. SSRI Research Report 78-7. Los Angeles: Social Science Research Institute, University of Southern California, 1978b.

Keeney, R. L., & Raiffa, H. *Decisions with multiple objectives*. New York: John Wiley, 1976.

Linstone, H. A., & Turoff, M. *The Delphi Method: Techniques and applications*. Reading, MA: Addison-Wesley, 1975.

McClelland, G. H. *Equal versus differential weighting for multiattribute decisions: There are no free lunches*. Report No. 207. Boulder, CO: Center for Research on Judgment and Policy, University of Colorado, 1978.

Newman, J. R. Differential weighting in multi-attribute utility measurement: When it should not and when it does make a difference. *Organizational Behavior and Human Performance*, 1977, *20*, 312-324.

Pitz, G., Sach, N., & Heirboth, J. Procedures for eliciting choices in the analysis of individual decisions. *Organizational Behavior and Human Performance*, 1980, *26*, 396-408.

Raiffa, H. *Performances for Multi-Attributed Alternatives*. RM-5868-DOT/RC. Santa Monica, CA: The Rand Corporation, 1969.

Rohrbaugh, J. Improving the quality of group judgment: Social judgment analysis and the Delphi Technique. *Organizational Behavior and Human Performance*, 1979, *24*, 73-92.

Rohrbaugh, J. Improving the quality of group judgment: Social judgment analysis and the nominal group technique. *Organizational Behavior and Human Performance*, 1981, *28*, 272-288.

Stillwell, W. G., Barron, F. H., & Edwards, W. *Evaluating credit applications: A validation of multiattribute utility techniques against a real world criterion*. Technical Report 80-1. Los Angeles: Social Science Research Institute, University of Southern California, 1980.

Stillwell, W. G., Seaver, D. A., & Edwards, W. A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior and Human Performance*, 1981, *28*, 62-77.

Stumpf, S. A., Freedman, R. D., & Zand, D. E. Judgment decisions: A study of interactions among group membership, group functioning, and the decision situation. *Academy of Management Journal*, 1979, *22*, 765-782.

von Winterfeldt, D., & Edwards, W. *Error in decision analysis: How to create the possibility of large losses by using dominated strategies*. SSRI Research Report 75-4. Los Angeles: Social Science Research Institute, University of Southern California, 1975.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, *83*, 213-217.

Winer, B. J. *Statistical principles in experimental design*. 2nd ed. New York: McGraw-Hill, 1971.

APPENDIX A

ANALYSTS' INSTRUCTIONS FOR EACH STRUCTURING TECHNIQUE: EXPERIMENT 1

TOP-DOWN STRUCTURING INSTRUCTIONS

Your role will be to help the participating second lieutenants construct a MAVA hierarchy for evaluating a company's performance in attacking the enemy during a combat readiness evaluation. The MAVA hierarchy must be built in a hierarchical fashion from the top-level node, which in this case is the overall value on ATTACK. All bottom-level nodes (i.e., attributes) must represent activities that can be scored "Pass or Fail" to measure a company's performance in attacking the enemy during an evaluation. These activities must be unique to the ATTACK Mission Performance Standard (MPS) in the Marine Corps Combat Readiness Evaluation System (MCCRES).

The ATTACK represents those specific actions required to make the final assault in daylight. It begins with planning activities and concludes with the unit's employment of reserves, response to counterattack, and command displacement. It should not include actions involving general command and control, fire support coordination, or movement to contact since these general categories represent separate mission performance standards (MPS) in MCCRES. Your structure, however, may include command and control, fire support coordination, and movement to contact activities that are unique to the ATTACK MPS. In addition, it may include actions by individual Marines that are unique to the ATTACK MPS. However, try not to emphasize these four classes of activities, so that your group has sufficient time to focus on activities clearly unique to attacking an enemy.

The following general steps are to be used by the analyst.

First, the analyst gets the group members to agree on a set of top-level categories. The set should be comprehensive so that it can subsume all action outcomes for measuring a USMC company's performance in attacking the enemy. But do not belabor the first draft. If later they suggest new additions or change at the top level, fine.

Second, the group is asked to decompose top-level categories into second-level categories, then third level, and so on until the group is satisfied that the categories cannot usefully be subdivided further. The number of nodes and levels may vary from one part of the tree to another.

In decomposing to the second level and beyond, you may ask the group to decompose each branch to the bottom before moving horizontally to the next node, or you may have them specify all nodes at a level before proceeding downward, or any combination which seems to work best. Be guided by the group's preference, but, generally, build from the top down rather than from the bottom up. That is, any branch in the completed structure should have had its top level defined before its second level, its second level before its third level, and so on. Note down any premature bottom-level suggestions for inclusion later, and accept all revisions they want in the structure, in whatever order they are made. All bottom-level nodes must represent specific activities (or outcomes) that can be used to measure whether or not the company successfully performed the actions involved in attacking the enemy. Decomposition continues until group members think they have defined all such activities explicitly.

Finally, the analysts must review the hierarchy with group members. This should be done from the top-down in the same manner the hierarchy was constructed. The analyst should encourage the group to modify the hierarchy as needed.

The analyst must photograph or hand-copy the hierarchy after the review process is completed. This will be our only record of the group's hierarchy. The hierarchy can be copied while the group is using the different weighting techniques.

Since each analyst may lead more than one session, the analyst should not offer substantive advice to the group members. For example, the analyst is never to tell the second lieutenants what attributes should be in the hierarchy. The analyst is to serve primarily as a methodologist whose job is to help group members structure the MAV hierarchy. Any substantive advice you offer should be as a result only of your interaction with the particular group of second lieutenants that you are working with during the session.

BOTTOM-UP STRUCTURING INSTRUCTIONS

Your role will be to help the participating second lieutenants construct a MAVA hierarchy for evaluating a company's performance in attacking the enemy during a combat readiness evaluation. The MAVA hierarchy must be built in a bottom-up fashion. All bottom-up level nodes (i.e., attributes) must represent activities that can be scored "Pass or Fail" to measure a company's performance in attacking the enemy during an evaluation. These activities must be unique to the ATTACK Mission Performance Standard (MPS) in the Marine Corps Combat Readiness Evaluation System (MCCRES).

The ATTACK MPS represents those specific actions required to make the final assault in daylight. It begins with planning activities and concludes with the unit's employment of reserves, response to a counterattack, and command displacement. It should not include actions involving general command and control, or fire support coordination, or movement to contact, because these general categories represent separate mission performance standards (MPS) in MCCRES. Your structure, however, may include command and control, fire-support coordination, and movement to contact activities that are unique to the attack MPS. In addition, it may include actions by individual Marines that are unique to the ATTACK MPS. Try not to emphasize these four classes of activities, however, so that your group has sufficient time to focus on activities clearly unique to attacking an enemy.

The following general steps are to be employed when using the bottom-up structuring approach to develop a hierarchy for measuring a company's performance in attacking an enemy.

First, get group members to identify all activities (or performance outcomes) they would want to measure in order to score a company's performance in attacking an enemy. That is, the analyst should get group members to identify all bottom-level nodes in the hierarchy. The group members may not always do so, particularly at the beginning of the session. Analysts may write intermediate-level nodes on the board for future reference, but they should continuously focus the group on the hierarchy's bottom-level nodes.

One way to get the group started in identifying bottom-level nodes is to ask the participants to describe the activities that would distinguish a good company from a bad company during an attack. These activities may be outcomes, as well as process activities. The first step continues until group members list all bottom-level nodes for measuring a company's ability to attack. Remember, bottom-level attributes represent activities that can be scored "pass" or "fail" during an evaluation.

Second, the analyst should get the group to cluster bottom-level nodes together in order to represent intermediate-level attributes in the hierarchy. The analysts may remind the group of previously identified intermediate-level attributes. However, the analyst should minimize substantive advice, such as telling the second lieutenants which attributes to cluster together. Any substantive advice should be as a result only of your interaction with the particular group of second lieutenants with whom you are working, since each analyst may lead more than one session. The group may have as many levels of intermediate attributes (and lower-level attributes per level) as they consider necessary.

Third, the analyst must review the MAV hierarchy with group members. This should be done from the bottom-up. The analyst may permit the group to modify the hierarchy.

The analyst must photograph or hand-copy the hierarchy after the review process is completed. This will be our only record of the group's hierarchy. The hierarchy can be copied while the group is using the different weighting techniques.

APPENDIX B

THE INSTRUCTIONS FOR EACH WEIGHTING TECHNIQUE, FOR THE FLAT, FIVE-ATTRIBUTE TASK: EXPERIMENT 2

INSTRUCTIONS

You will now be asked to indicate the relative importance of getting a "Yes" on each of the five requirements for performing the task "PRELIMINARY OPERATIONS" in the ATTACK MPS. The five requirements are:

A. Patroling intensified
B. Reconnaissance elements dispatched
C. Indirect fire emplaced well forward
D. Harassing fires delivered
E. Assault elements moved into attack position

(See accompanying page for a full description.)

Please use the steps on the following pages to indicate the relative importance of succeeding in each requirement. Use the attached form for your answers.

1. Rank-order the requirements in terms of the relative importance of performing them successfully. The requirement that is most important to perform successfully in order to perform well on the "PRELIMINARY OPERATIONS" task should be ranked #1. The next most important requirement to perform successfully should be ranked #2, and so forth. You are allowed to have "ties," indicating that two requirements are equally important.

2. Assign a score of 10 to the least important requirement to perform successfully; this is the fifth most important requirement (ranked #5). All other requirements will be scored in relation to the 10 points given the fifth most important requirement.

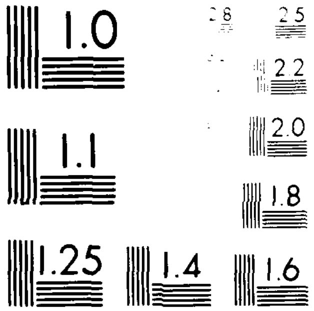
3. Assign a score to the fourth most important requirement to perform successfully (ranked #4). This score must be equal to or greater than 10. This score is a ratio score. So, if successfully performing the fourth most important requirement (#4) is twice as important as successfully performing the fifth most important requirement (#5), it should get a 20. If it is ten times as important, it should get a 100. If it is just as important, it should get a 10. If it is only half again as important as #5, it should receive a 15. If it is four and one-half times as important, it should receive a score of 45.

4. Assign a score to the third most important requirement (#3) to perform successfully. The score must be equal to or greater than that given the fourth most important requirement (#4). For example, if the fourth most important requirement received a score of 50, the score on the third most important requirement must be equal to or greater than 50. Furthermore, its score must be

AD-A111 543   DECISIONS AND DESIGNS INC MCLEAN VA   F/G 12/1
EVALUATING THE RELATIVE EFFECTIVENESS OF DIFFERENT STRUCTURING --ETC(U)
JAN 82 L ADELMAN, P J STICHA, M L DONNELL   N00014-81-C-0022
UNCLASSIFIED   TR-82-1-326-13   NL

2 OF 2
AD A
1 543

END
DATE
FILMED
03-82
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

relative to that given the fourth and fifth most important requirements. If, in our example, the third requirement is twice as important as #4, it should receive a score of 100 because #4 received a score of 50. This implies that it is ten times as important as the fifth most important requirement (#5). If this is not true, then the score given the third requirement or that given the fourth requirement must be changed. In our example, assume that performing the third most important requirement is not ten times, but five times as important as than fifth most important requirement (#5). Then, the score on the fourth requirement must be changed so that it is between 10 and 50. If we now assume that #4 is seven-eighths of the distance between 10 and 50, it should be given a score of 45 since $10 + \frac{7}{8} \times 40 = 10 + 35 = 45$. This means that requirement #4 is 4 1/2 times as important as requirement #5.

Notice that in our example, the scores given the fifth and fourth most important requirements add up to more than 50, since 10 + 45 = 55. This means that you would prefer to perform the fifth and fourth requirements successfully and fail the third one rather than perform the third requirement successfully and fail the fourth and fifth requirements. If this were not true, you would have to either lower the scores on the fourth requirement or raise the score on the third requirement so that the fourth and fifth most important requirements added up to less than the third most important requirement.

5. Assign a score to each of the other requirements to indicate the importance of performing it successfully relative to the other requirements. Make sure to check the implications of your scores by adding them together, as we illustrated in step 4. For example, is it less

important to perform requirements #3, #4, and #5 successfully and fail requirement #2 or pass requirement #2 and fail #3, #4 and #5? If more important to pass #2, how much more important is it?

ROOM ____________ PARTICIPANT NUMBER ____________

DATE ____________ TASK ____________

ANSWER SHEET

	RANK-ORDER	SCORES
	List the Requirements From the Most to Least Important One to Perform Successfully	Assign Relative Scores to Each Requirement
#1		
#2		
#3		
#4		
#5		10

INSTRUCTIONS

On the following pages, you will be asked a number of "paired comparison" questions for the following five requirements for the task "PRELIMINARY OPERATIONS."

A. Patroling intensified

B. Reconnaissance elements dispatched

C. Indirect fire emplaced well forward

D. Harassing fires delivered

E. Assault elements moved into attack position

The first thing you must do to answer these questions is rank the five requirements in the order of their importance. Use the evaluation form on the last page. The most important requirement should be ranked #1, the second most important requirement should be ranked #2, and so forth until all five requirements have been ranked from most to least important in performing PRELIMINARY OPERATIONS.

After rank-ordering the five requirements, answer the paired-comparison questions on the following pages. The first comparison is between the requirement you ranked as most important (#1) and the two next most important requirements (#2 and #3). You are to indicate whether the successful performance of #1 is

(a) much more important (b) more important (c) as important (d) less important or (e) much less important

than the successful performance of both requirements #2 and #3. Write the letter that you consider true. For example, you would write (c) if you thought that the successful performance of requirement #1 was just as important as the successful performance of both requirements #2 and #3.

You would write (a) if you thought that the successful performance of requirement #1 was much more important than the successful performance of both requirements #2 and #3.

You will have to answer eleven paired-comparison questions. Answer each question by selecting one of the following five choices:

(a) much more important (b) more important (c) as important (d) less important (e) much less important

Write your answer to each paired comparison question on the answer sheet. Make sure your answers are consistent. For example, if you say that requirement #2 is much more important than requirements #3, #4, and #5, then requirement #2 also must be much more important than requirements #3 and #4 or requirements #3 and #5 or requirements #4 and #5. Remember, rank-order the requirements in the order of their importance before you begin answering the paired-comparison questions.

After you have answered the paired comparison questions, you must indicate the importance of the requirement you ranked #4 relative to the requirement you ranked #5. Arbitrarily assign a score of 10 to the fifth most important requirement to perform successfully (#5). Now, assign a score to the fourth most important requirement to perform successfully (ranked #4). This score must be equal to or greater than 10. This score is a ratio score. So, if successfully performing #4 is twice as important as successfully performing #5 it should get a 20. If it is ten times as important, it should get a 100. If it is just as important, it should get a 10. If #4 is only half again as important as #5, it should receive a 15. If it is four and one-half times as important, it should receive a score of 45. Write your score for requirement #4 in the space provided on the answer sheet.

ROOM__________ PARTICIPANT NUMBER__________

ANSWER SHEET

RANK-ORDER

List the Requirements From the
Most to Least Important One to
Perform Successfully

#1_________________________

#2_________________________

#3_________________________

#4_________________________

#5_________________________

PAIRED-COMPARISONS

Answer each by selecting one of the following five choices:

(a) much more important (b) more important (c) as important (d) less important (e) much less important

1. #1 vs #2 and #3? __________

2. #1 vs #3 and #4? __________

3. #1 vs #4 and #5? __________

4. #1 vs #2, #3, and #4? __________

5. #1 vs #3, #4, and #5? __________

6. #1 vs #2, #3, #4, #5? __________

7. #2 vs #3 and #4? __________

8. #2 vs #4 and #5? __________

9. #2 vs #3 and #5? __________

10. #2 vs #3, #4, and #5? __________

11. #3 vs #4 and #5? __________

If #5 was given a score of 10, what would be your score for #4? __________

KRH

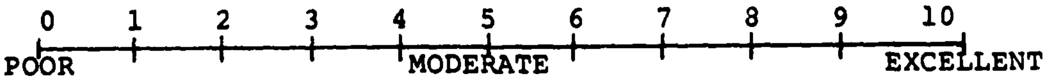
INSTRUCTIONS

You will now be asked to evaluate how well different battalions have performed the task PRELIMINARY OPERATIONS in the "ATTACK" Mission Performance Standard. Each battalion's evaluation will depend on its success on the five requirements for performing the task. For example, here is a hypothetical battalion's performance on the five requirements; P stands for "Passed" and F stands for "Failed."

Patroling intensified	P
Reconnaissance elements dispatched	F
Indirect fire emplaced well forward	P
Harassing fires delivered	P
Assault elements moved into attack position	P

As you can see, the battalion passed all of the requirements except dispatching reconnaissance elements. This failure does not mean that reconnaissance elements were not dispatched, but that they were not dispatched according to specified standards. (Note: In all cases, a failure means that the requirement was performed, but not according to specified standards. Rely on your experience in defining "failure" as the average or general level of poor performance on the requirement.)

Now, on the basis of this performance, we want you to give the battalion an overall score on "PRELIMINARY OPERATIONS" that indicates how well you think the battalion performed that task. Use a 0 to 10 scale to score the battalion where 0 is a poor score, and 10 is an excellent score, and 5 is a moderate score. The rating score is shown below.



Battalions that, in your opinion, performed well should receive higher scores than those that performed poorly. For example, battalions that failed all requirements, should receive a score of 0. Battalions that passed all requirements should receive a score of 10. Battalions that passed one or more requirements can receive any score on the 0 to 10 rating scale. Feel free to use fractions (e.g., 2.2 or 7.5) if you want to do so.

You will have to evaluate 23 battalions on the following pages. Each battalion is described in terms of its performance on the five requirements for performing the task "PRELIMINARY OPERATIONS." Remember, a failure means that the requirement was performed, but not according to specified standards. In all cases, your scores should reflect your intuitive judgment about how well the battalion performed based on the pattern of passes and fails. The first three battalions are for practice. The analyst working with you will lead you through them.

Use the answer sheets to score the 23 battalions. The first answer sheet is for the practice battalions; the second answer sheet is for the other twenty battalions. You may change the score for any battalion, if you choose to do so.

PRACTICE BATTALIONS

Practice Battalion #1

Patroling intensified	F
Reconnaissance elements dispatched	F
Indirect fire emplaced well forward	F
Harassing fires delivered	F
Assault elements moved into attack position	F

Practice Battalion #2

Patroling intensified	P
Reconnaissance elements dispatched	P
Indirect fire emplaced well forward	P
Harassing fires delivered	P
Assault elements moved into attack position	P

Practice Battalion #3

Patroling intensified	P
Reconnaissance elements dispatched	P
Indirect fire emplaced well forward	P
Harassing fires delivered	F
Assault elements moved into attack position	F

ROOM ________ PARTICIPANT NUMBER ________

EVALUATION SHEET

PRACTICE BATTALION #1 ____________

PRACTICE BATTALION #2 ____________

PRACTICE BATTALION #3 ____________

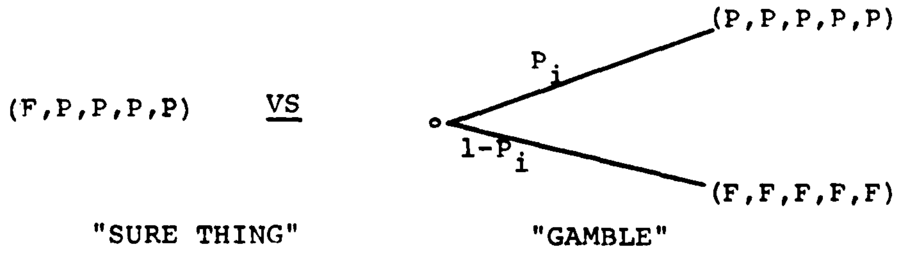
INSTRUCTIONS

Uncertainty about future performance has considerable implications for choosing battalions for different assignments. We want you to help us systematically evaluate this uncertainty with regard to a battalion's performance on the requirements for the task "PRELIMINARY OPERATIONS." The five requirements are:

1. Patrol intensified
2. Reconnaissance elements dispatched
3. Indirect fire emplaced well forward
4. Harassing fires delivered
5. Assault elements moved into attack position

(See accompanying page for a full description)

You will be asked to make five comparisons between a sure thing and a gamble. For example, the first comparison is between (a) a battalion failing requirement #1 (Patroling Intensified) and passing all other requirements for sure, and (b) a gamble where the battalion has a probability (p_i) of passing all of the requirements and a probability $(1-p_i)$ of failing all the requirements. This choice is shown in the figure below.



We want you to specify the value of p_i that makes you indifferent between accepting the gamble and getting the sure thing.

A good way to think about the choice is as follows. Assume that you had an urn filled with 1000 balls. A certain percentage (p_i) of the balls are white to represent the battalion in the gamble that passed all the requirements, and a certain percentage ($1-p_i$) of the balls are black to represent the battalion that failed all the requirements. What percentage (p_i) of white balls would the urn have to contain in order for you to be indifferent to receiving the "SURE THING" or the gamble?

For example, what would be your preference if the urn had 999 white balls and 1 black ball? Well, you would probably prefer the gamble to the "SURE THING" because you have a probability of .999 (out of a total of 1.0) of getting the battalion that passed all the requirements and a probability ($1-p_i$) of .001 of getting that battalion that failed all the requirements. In contrast, you would probably prefer the "SURE THING" to the gamble if the urn had 1 white ball (i.e., $p_i = .001$) and 999 black balls (i.e., $1-p_i = .999$). What if the urn had 500 white balls ($p_i = .5$) and 500 black balls ($1-p_i = .5$)? Would you prefer the "SURE THING" (i.e., the battalion that failed "Patroling intensified" and passed the other requirements) or the gamble (i.e., a 50% chance of getting the battalion that passed all the requirements and a 50% chance of getting the battalion that failed all the requirements)? Your job is to indicate the percentage of white balls (p_i) that would make you indifferent between receiving the "SURE THING" and receiving the gamble.

When making your choice, only consider the implications for readiness. Do not consider personal consequences of the utilization of the unit in a combat situation. In answering these comparisons, for example, don't worry abcut being relieved if your unit were to fail everything; instead, consider the implications for the readiness of that unit.

In all of the comparisons the battalion will have failed only one of the requirements in the sure thing. The gamble will always involve passing all the requirements with some value p_i versus failing all requirements with the probability $1-p_i$. In all cases, a failure means that the requirement was performed, but not according to specified standards. Your job is to indicate the value of p_i such that you would be indifferent between playing the gamble and having the sure thing.

All of the comparisons are listed in the next page. Below each comparison, first indicate if p_i is closest to 1.0, .75, .50, .25, or 0.0 by circling the appropriate letter. Then, indicate the actual p_i value such that you would be indifferent between receiving the sure thing or playing the gamble.

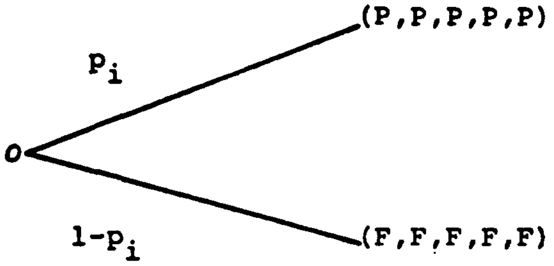
TASK __________ DATE __________

ROOM __________ PARTICIPANT'S NUMBER __________

ANSWER SHEET

Compare each of the five sure-things to the following gamble:

"SURE-THING" vs



1. (F,P,P,P,P) vs gamble
   - o is p_i closest to (circle one)
     (a) 1.0 (b) .75 (c) .50 (d) .25 (e) 0.0
   - o p_i is __________

2. (P,F,P,P,P) vs gamble
   - o Is p_i closest to (circle one)
     (a) 1.0 (b) .75 (c) .50 (d) .25 (e) 0.0
   - o p_i is __________

3. (P,P,F,P,P) vs gamble
   - o Is p_i closest to (circle one)
     (a) 1.0 (b) .75 (c) .50 (d) .25 (e) 0.0
   - o p_i is __________

4. (P,P,P,F,P) vs gamble
   - o Is p_i closest to (circle one)
     (a) 1.0 (b) .75 (c) .50 (d) .25 (e) 0.0
   - o p_i is __________

5. (P,P,P,P,F) vs gamble
   - o Is p_i closest to (circle one)
     (a) 1.0 (b) .75 (c) .50 (d) .25 (e) 0.0
   - o p_i is __________

ROOM __________ PARTICIPANT NUMBER __________

INSTRUCTIONS

At this time we would like you to indicate the relative importance of each requirement in performing the task, PRELIMINARY OPERATIONS. To accomplish this, please divide up 100 points between the requirements. The more important you consider the requirement, the higher its score. Please make sure that the points given to all of the requirements add up to 100 points.

Patroling intensified	__________
Reconnaissance elements dispatched	__________
Indirect fire emplaced well forward	__________
Harassing fires delivered	__________
Assault elements moved into attack position	__________
Sum =	100

ROOM ________ PARTICIPANT NUMBER ________

RELATIVE WEIGHTS ASSIGNED BY EACH OF THE TECHNIQUES
(SUM TO 100)

REQUIREMENTS	RANKING & RATING	PAIRED COMPARISONS	BATTALION EVALUATION	GAMBLE	DIVIDING UP 100 PTS.	FINAL* WEIGHTS
1.						
2.						
3.						
4.						
5.						

*Please indicate the relative weights you want to use to indicate the relative importance of the requirements. These weights may be the same as those generated by one of the five techniques or different if you think no technique accurately reflected your opinion. Write the weights in the column under "Final Weights."

APPENDIX C

THE ANALYSTS' INSTRUCTIONS FOR EACH DISCUSSION TECHNIQUE: EXPERIMENT 3

NOMINAL GROUP TECHNIQUE

The Nominal Group Technique (NGT) will be used to obtain your group's weights for the requirements (bottom-level nodes) for each of the four tasks. This technique has the following steps:

1. Each participant uses the specified weighting technique to obtain a set of weights for the requirements.

2. In round-robin fashion, each participant informs the group of his weights and the reasons for them. The analyst writes the weights and reasons on the board. These weights should not be normalized; they represent the values generated by the specified technique. Other group members are not permitted to argue with the participant about whether or not they consider the weights to be correct. They may, however, ask the participant to clarify his explanation of his weights.

3. After each participant has informed the group of his weights and the reasons for them, group members are to discuss areas of disagreement. During the group's discussion, the analyst should help group members to adhere to the following six guidelines: (1) avoid arguing; (2) avoid "I'm right, you're wrong" statements; (3) avoid changing their opinions only to avoid conflict and to reach agreement and harmony; (4) avoid conflict-reducing techniques such as the majority vote, averaging, bargaining, coin flipping; (5) view differences of opinion as both natural and helpful rather than as a hindrance in decision making; and (6) view initial agreement as suspect. In short,

the discussion should be open, with each participant feeling free to discuss his position with others. The analyst's should feel free to pace the group so that group weights are obtained for all four tasks within the four-hour time limit. (Note: The analyst should record the start and finish time for each task; that is, from the beginning of Step 2 to the end of Step 3.)

4. When the analyst thinks the major points have been discussed by the group, the analyst should bring the group's discussion to an end. Using the specified weighting technique, each participant should again individually specify what he considers to be the correct weights.

5. The analyst collects the final weights. Have group members complete their rating forms. In addition, complete your rating form, too.

UNSTRUCTURED GROUP DISCUSSION TECHNIQUE

An unstructured group discussion technique will be used to obtain your group's weights for the requirements (bottom-level nodes) for each of the four tasks. This technique includes the following steps:

1. Each participant uses the specified weighting technique to obtain a set of weights for the requirements. These weights are primarily for subsequent analysis; therefore, do not encourage the group to focus on each individual member's weights. Collect each participant's data sheet before beginning the group discussion.

2. Have the group specify the weights for the requirements. It is the analyst's job to lead the group's discussion; meaning that group members interact primarily with the analyst. This is different than the Nominal Group Technique where group members interact primarily with each other, not the analyst.

   During the group's discussion, the analyst should help group members to adhere to the following six guidelines: (1) avoid arguing; (2) avoid "I'm right, you're wrong" statements; (3) avoid changing their opinions only to avoid conflict and to reach agreement and harmony; (4) avoid conflict-reducing techniques such as the majority vote, averaging, bargaining, coin flipping; (5) view differences of opinion as both natural and helpful rather than as a hindrance in decision making; and (6) view initial agreement as suspect. In short, the discussion should be open, with each participant feeling free to discuss his position with others.

The weights should not be normalized yet; instead, they represent the values generated by the specified technique. The analyst should feel free to pace the group so that group weights are obtained for all four tasks within the four-hour time limit. (Note: The analyst should record the start and finish time for each task; that is, from the beginning of Step 2 to the end of Step 3.)

3. After the group agrees on their weights, each participant individually specifies what he considers to be the correct weights. This final set of weights may be the same or different from the group's weights.

4. Have group members complete their rating form. In addition, complete your rating form, too.

5. Normalize the group's weights so that the weights sum to 100. Record the normalized weights on the form titled "Group Weights".

OFFICE OF NAVAL RESEARCH

TECHNICAL REPORTS DISTRIBUTION LIST

Engineering Psychology Programs
Code 442
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217 (5 cys)

Operations Research Programs
Code 411-OR
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Statistics and Probability Program
Code 411-S&P
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Information Systems Program
Code 411-IS
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

CDR K. Hull
Code 410B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Physiology & Neuro Biology Programs
Code 441B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Commanding Officer
ONR Eastern/Central Regional Office
ATTN: Dr. J. Lester
Building 114, Section D
666 Summer Street
Boston, MA 02210

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, DC 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, DC 20375

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, DC 20350

Human Factors Department
Code N215
Naval Training Equipment Center
Orlando, FL 32813

Dr. Alfred F. Smode
Training Analysis and Evaluation
Group
Naval Training Equipment Center
Code N-00T
Orlando, FL 32813

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Mr. Warren Lewis
Human Engineering Branch
Code 8231
Naval Ocean Systems Center
San Diego, CA 92152

Commanding Officer
ONR Western Regional Office
ATTN: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, DC 20380

Mr. Arnold Rubinstein
Naval Material Command
NAVMAT 0722 - Rm. 508
800 North Quincy Street
Arlington, VA 22217

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 340F
Washington, DC 20361

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Center
Bethesda, MD 20014

Dr. Arthur Bachrach
Behavioral Sciences Department
Naval Medical Research Institute
Bethesda, MD 20014

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Head
Aerospace Psychology Department
Code L5
Naval Aerospace Medical Research Lab
Pensacola, FL 32508

Dr. James McGrath
CINCLANT FLT HQS
Code 04E1
Norfolk, VA 23511

Dr. Robert Blanchard
Navy Personnel Research and
Development Center
Command and Support Systems
San Diego, CA 92152

LCDR Stephen D. Harris
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Dr. Julie Hopson
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons Center
China Lake, CA 93555

Human Factors Engineering Branch
Code 1226
Pacific Missile Test Center
Point Mugu, CA 93042

CDR W. Moroney
Code 55MP
Naval Postgraduate School
Monterey, CA 93940

Dr. Joseph Zeidner
Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and
Systems Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, DC 20332

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Navy Personnel Research and
Development Center
Planning & Appraisal Division
San Diego, CA 92152

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology
Establishment
Teddington, Middlesex TW11 OLN
ENGLAND

Director, Human Factors Wing
Defence & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9
CANADA

Dr. A. D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
ENGLAND

Defense Technical Information Center
Cameron Station, Bldg. 5
Alexandria, VA 22314

Dr. Judith A. Daly
Defense Advanced Research Projects
Agency
DSO/SSD
1400 Wilson Blvd.
Arlington, VA 22209

Dr. Robert R. Mackie
Human Factors Research, Inc.
5775 Dawson Avenue
Goleta, CA 93017

Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, VA 22202

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
455 West Lindsey
Norman, OK 73069

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks AFB, TX 78235

Dr. T. B. Sheridan
Department of Mechanical
Engineering
Massachusetts Institute of
Technology
Cambridge, MA 02139

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Dr. Robert T. Hennessy
NAS - National Research Council
2101 Constitution Ave., N.W.
Washington, DC 20418

Dr. Michael G. Samet
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

Dr. Robert Williges
Human Factors Laboratory
Virginia Polytechnic Institute
and State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Gary McClelland
Institute of Behavioral Sciences
University of Colorado
Boulder, CO 80309

Dr. Ward Edwards
Director, Social Science
Research Institute
University of Southern
California
Los Angeles, CA 90007

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, CO 80309

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, DC 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Dr. Christopher Wickens
University of Illinois
Department of Psychology
Urbana, IL 61801

Dr. Richard W. Pew
Information Sciences Division
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

Dr. Hillel Einhorn
University of Chicago
Graduate School of Business
1101 E. 58th Street
Chicago, IL 60637

Dr. John Payne
Duke University
Graduate School of Business
Administration
Durham, NC 27706

Dr. Baruch Fischhoff
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Andrew P. Sage
University of Virginia
School of Engineering and Applied
Science
Charlotesville, VA 22901

Dr. Lola Lopes
Department of Psychology
University of Wisconsin
Madison, WI 53706

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Industrial Park
Burlington, MA 01803

Dr. Rex Brown
Decision Science Consortium
Suite 721
7700 Leesburg Pike
Falls Church, VA 22043

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090

Additional Distribution

LCOL Paul R. Catalogne
Industrial College of the Armed
Forces
Fort Leslie McNair
Washington, DC 20319

Major Ray McCormick
Chief, TOC Group
The Basic School
Marine Corps Development and
Education Command
Quantico, VA 22134

LCOL Ron Roth
Operations Division
Plans, Policies, and Operations
Department
Headquarters, U.S. Marine Corps
(Code OTOR)
Washington, DC 20360

DAT

FILM